1 An Introduction to GIS

(Geographical Information System)

Introduction

Geography has always been important to humans. Stone-age hunters anticipated the location of their quarry, early explorers lived or died by their knowledge of geography, and current societies work and play based on their understanding of who belongs where. Applied geography, in the form of maps and spatial information, has served discovery, planning, cooperation, and conflict for at least the past 3,000 years (Figure 1-1). Maps are among the most beautiful and useful documents of human civilization, and spatial information has a great impact on our lives by helping us produce the food we eat, the energy we burn, the clothes we wear, and the diversions we enjoy.

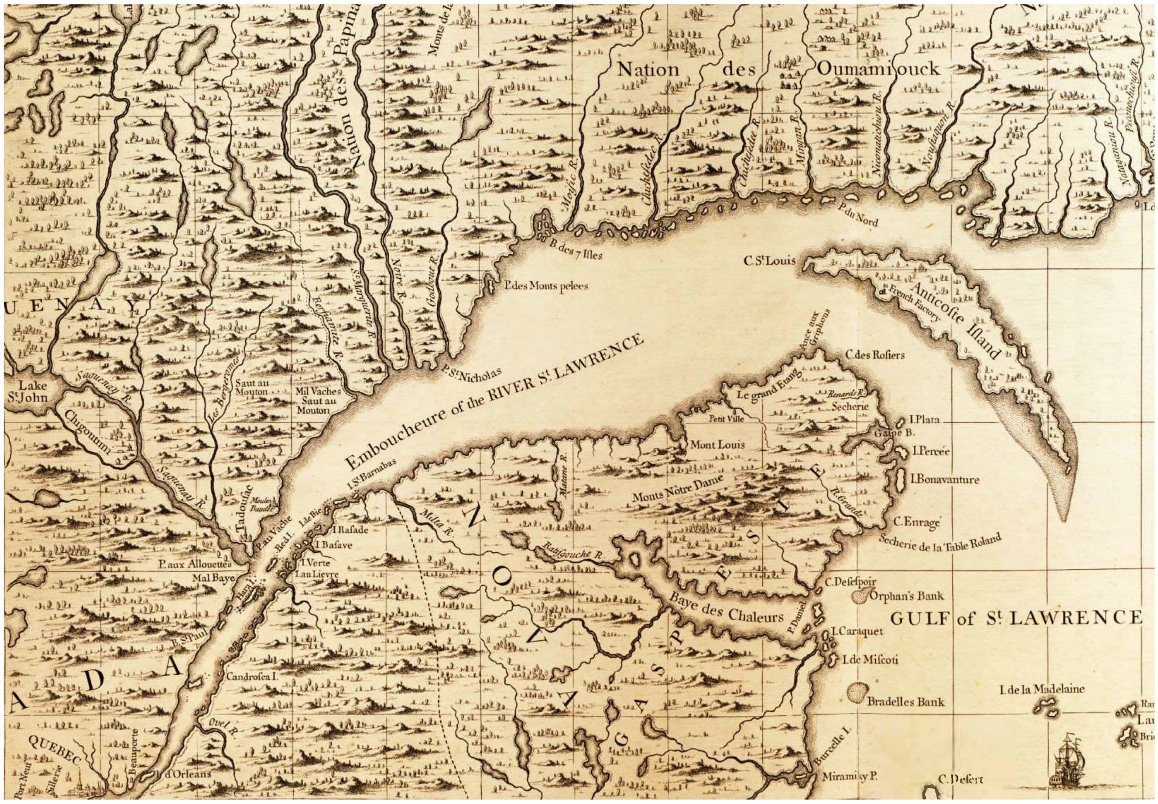


Figure 1-1: A map of the mouth of the St. Lawrence River, most probably by Clement Lempiere, published in 1733. The river mouth is in the center, New Brunswick lower center, and Quebec across the top. Early maps were key to exploration (courtesy U.S. Library of Congress).

Because spatial information is so important, we have developed tools called geographic information systems (GIS) to aid us with geographic knowledge. A GIS helps us gather and use spatial data (we will use the abbreviation GIS to refer to both singular, system, and plural, systems). Some GIS components are purely technological; these include space-age data collectors, advanced communications networks, and sophisticated computing. Other GIS components are very simple, for example, a pencil and paper used to field-verify a map.

As with many aspects of life in the last five decades, how we gather and use spatial data has been profoundly altered by modern electronics, and GIS software and hardware are primary examples of these technological developments. The capture and analysis of spatial data has accelerated over the past four decades, and continues to evolve.

Key to all definitions of a GIS are "where" and "what." GIS record the absolute and relative location of features, as well as the properties and attributes of those features. Mount Everest is in Asia, Timbuktu is in Mali, and the cruise ship *Titanic* is at the bottom of the Atlantic Ocean. A GIS quantifies these locations by recording their *coordinates*, numbers that describe the position of these features on Earth. The GIS may also be used to record the height of Mount Everest, the population of Pierre, or the depth of the *Titanic*, as well as any other defining characteristics of each spatial feature.

What is a GIS?

A GIS is a tool for making and using spatial information. Among the many definitions of GIS, we choose:

> *A GIS is a computer-based system to aid in the collection, maintenance, storage, analysis, output, and distribution of spatial data and information.*

When used wisely, GIS can help us live healthier, wealthier, and safer lives.

Each GIS user may decide what features are important, and what attributes are worth recording. For example, forests are important to many of us. They may protect water supplies, yield wood, harbor wildlife, and provide space to recreate (Figure 1-2).

Figure 1-2: GIS allow us to analyze important geographic features. The satellite image at the center shows a forested area in western Oregon, with a patchwork of lakes, forests, clearings, alpine zones, and deserts. A GIS may aid in ensuring sustainable recreation, timber harvest, environmental protection, and other benefits (courtesy NASA).

We are concerned about the level of harvest, the adjacent land use, pollution from nearby industries, or where forests burn. Informed management requires knowledge of all these related factors and, perhaps above all, the spatial arrangement of these factors. Buffer strips near rivers may protect water supplies, clearings may prevent the spread of fire, and polluters downwind may not harm our forests while polluters upwind might. A GIS helps us analyze these spatial interactions, and is also particularly useful at displaying spatial data and analysis. A GIS is often the only way to solve spatially-related problems.

Why We Need GIS

GIS are essential tools in business, government, education, and nonprofit organizations, and GIS use has become mandatory in many settings. GIS have been used to fight crime, protect endangered species, reduce pollution, cope with natural disasters, treat epidemics, and improve public health; in short, GIS have been instrumental in addressing some of our most pressing societal problems.

GIS tools in aggregate save billions of dollars annually in the delivery of governmental and commercial goods and services. GIS regularly help in the day-to-day management of many natural and man-made resources, including sewer, water, power, and transportation networks. GIS are at the heart of one of the most important processes in U.S. democracy, the constitutionally mandated reshaping of U.S. congressional districts, and hence the distribution of tax dollars and other government resources.

GIS are needed in part because human populations and consumption have reached levels such that many resources, including air and land, are placing substantial limits on human action (Figure 1-3). Human populations have doubled in the last 50 years, surpassing 7 billion, and we will likely add another 4 billion humans in the next 50 years. The first 100,000 years of human existence caused scant impacts on the world's resources, but in the past 300 years humans have permanently altered most of

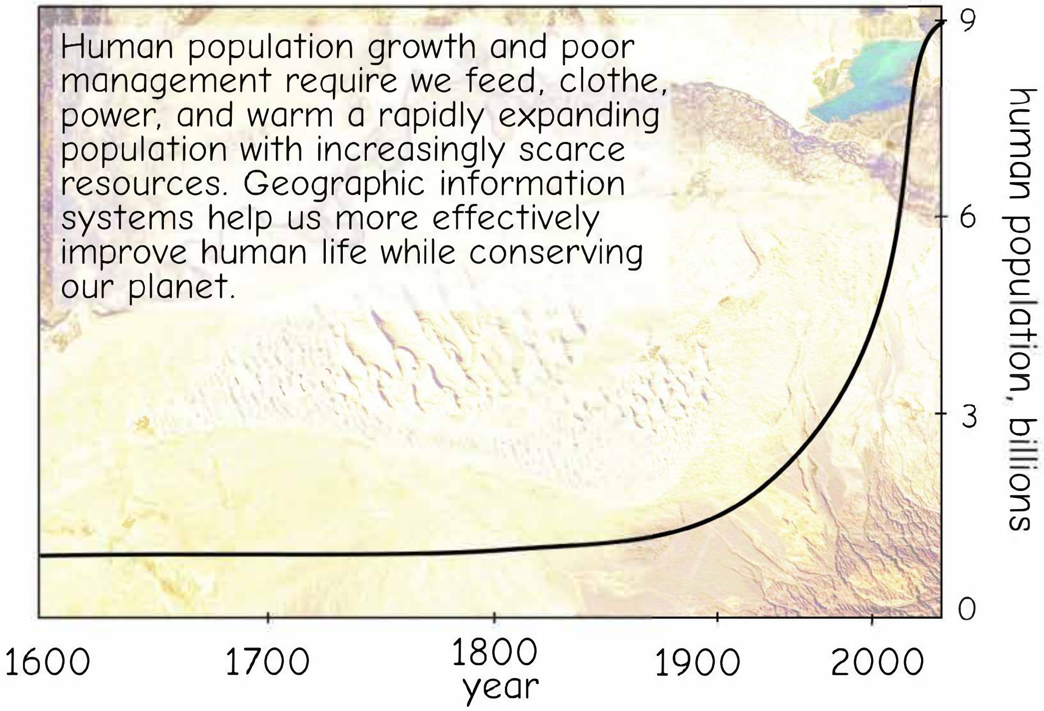


Figure 1-3: Human population growth during the past 400 years has increased the need for efficient resource use (courtesy United Nations and Ikonos).

the Earth's surface. The atmosphere and oceans exhibit a decreasing ability to benignly absorb carbon dioxide and nitrogen, two primary waste products of humanity. Silt chokes many rivers, and there are abundant examples of smoke, ozone, or other noxious pollutants substantially harming public health. By the end of the 20th century, most lands south of the boreal region had been farmed, grazed, cut, built over, drained, flooded, or otherwise altered by humans (Figure 1-4).

GIS help us identify and address environmental problems by providing crucial information on where problems occur and who are affected by them. GIS help us identify the source, location, and extent of adverse environmental impacts, and may help us devise practical plans for monitoring, managing, and mitigating environmental damage.

Human impacts on the environment have spurred a strong societal push for the adoption of GIS. Conflicts in resource use, concerns about pollution, and precautions to protect public health have led to legislative mandates that explicitly or implicitly require the consideration of geography. The U.S. Endangered Species Act of 1973 (ESA) is an example of the importance of geography in resource management. The ESA requires adequate protection of rare and threatened organisms. Effective protection entails mapping the available habitat and analyzing species range and migration patterns. The location of viable remnant plant and animal populations relative to current and future human land uses must be analyzed, and action taken to ensure species survival. GIS have proven to be useful tools in all of these tasks. GIS use is mandated in other endeavors, including emergency services, flood protection, disaster assessment and management (Figure 1-5), and infrastructure development.

Public organizations have adopted GIS because of legislative mandates, and because GIS aid in governmental functions. For example, emergency service vehicles are regularly dispatched and routed using GIS. E911 callers and addresses are automatically identified by telephone number. The GIS matches the address to the nearest emergency service station, a route is then immediately generated based on the street network and traffic, and emergency crews dispatched from the nearest station.

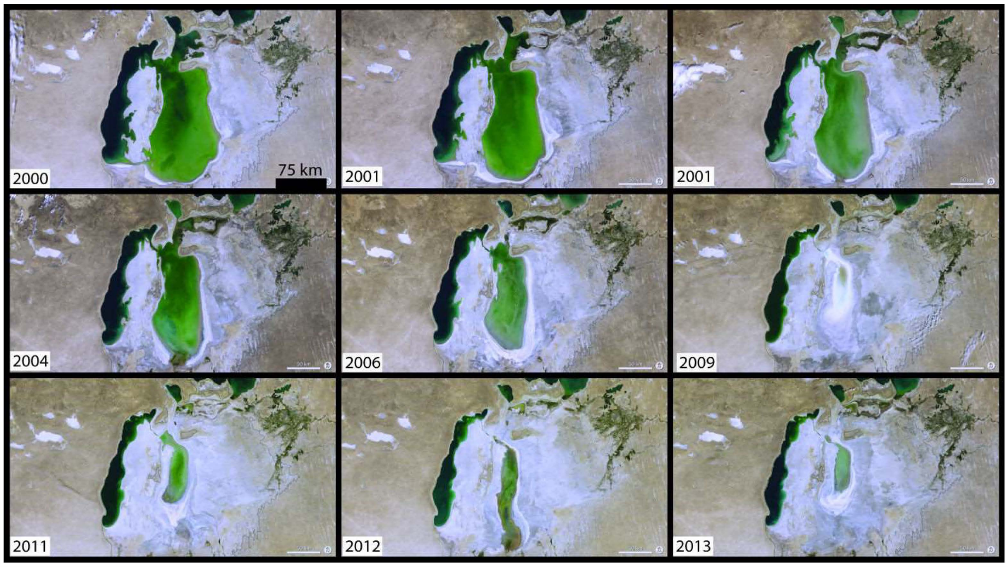


Figure 1-4: The environmental impacts wrought by humans have accelerated in many parts of the world during the past century. These satellite images from 2000 (upper left) to 2013 (lower right) show a shrunken Aral Sea due to the overuse of water. Diversion for irrigation has destroyed a rich fishery, the economic base for many seaside communities. GIS may be used to document change, mitigate damage, and effectively manage our natural resources (courtesy NASA).

Figure 1-5: GIS may aid in disaster assessment and recovery. These satellite images from Banda Aceh, Indonesia, illustrate tsunami-caused damage to a shoreline community. Emergency response and longer-term rebuilding efforts may be improved by spatial data collection and analysis (courtesy DigitalGlobe).

Many businesses adopt GIS for increased efficiency in the delivery of goods and services. Retail businesses locate stores based on a number of spatially related factors. Where are the potential customers? What is the spatial distribution of competing businesses? Where are potential new store locations? What are traffic flows near current stores, and how easy is it to park near and access these stores? GIS are also used in hundreds of other business applications, to route delivery vehicles, guide advertising, design buildings, plan construction, and sell real estate.

The societal push to adopt GIS has been complemented by a technological pull in the development and application of GIS. Thousands of lives and untold wealth have been lost because ship captains could not answer the simple question, “Where am I?” Robust nautical navigation methods emerged in the 18th century, and have continually improved since, so that anyone can quickly locate their outdoor position to within a few meters. Remarkable positioning technologies, generically known as Global Navigation Satellite Systems (GNSS), are now indispensable tools in commerce, planning, and safety.

The technological pull has developed on several fronts. Spatial analysis in particular has been helped by faster computers with more storage, and by the increased interconnectedness via WiFi and mobile networks. Most real-world spatial problems were beyond the scope of all but the largest government and business organizations until the 1990s. GIS computing expenses are becoming an afterthought, as computing resources often cost less than a few weeks’ salary for a qualified GIS professional.Costs decrease and performance increases at dizzying rates, with predicted plateaus pushed back each year. Powerful field computers are lighter, faster, more capable, and less expensive, so spatial data display and analysis capabilities may always be at hand (Figure 1-6).

GIS on rugged, field-portable computers has been particularly useful in field data entry and editing.

In addition to the computing improvements and the development of GNSS, current "cameras" deliver amazingly detailed aerial and satellite images. Initially, advances in image collection and interpretation were spurred by World War II and then the Cold War because accurate maps were required, but unavailable. Turned toward peacetime endeavors, imaging technologies now help us map food and fodder, houses and highways, and most other natural and human-built objects. Images may be rapidly converted to accurate spatial information over broad areas (Figure 1-7). Many techniques have been developed for extracting information from image data, and also for ensuring this information faithfully represents the location, shape, and characteristics of features on the ground. Visible light, laser, thermal, and radar scanners are currently being developed to further increase the speed and accuracy with which we map our world. Thus, advances in these three key technologies — imaging, GNSS, and computing — have substantially aided the development of GIS.

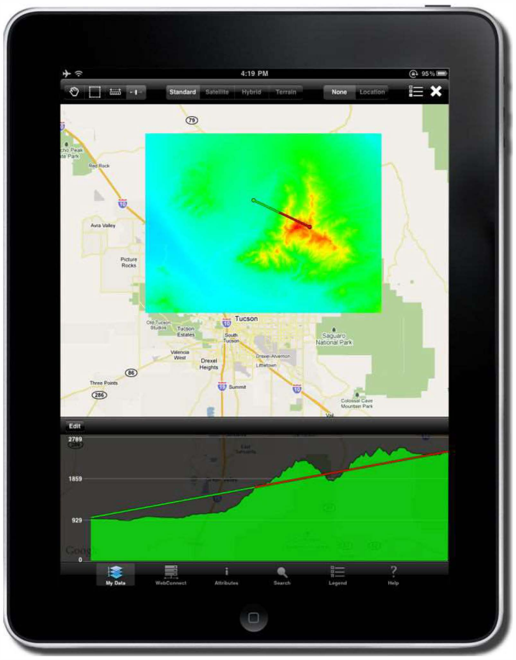


Figure 1-6: Portable computing is one example of the technological pull driving GIS adoption (courtesy Cogent3D, www.GISRoam.com).

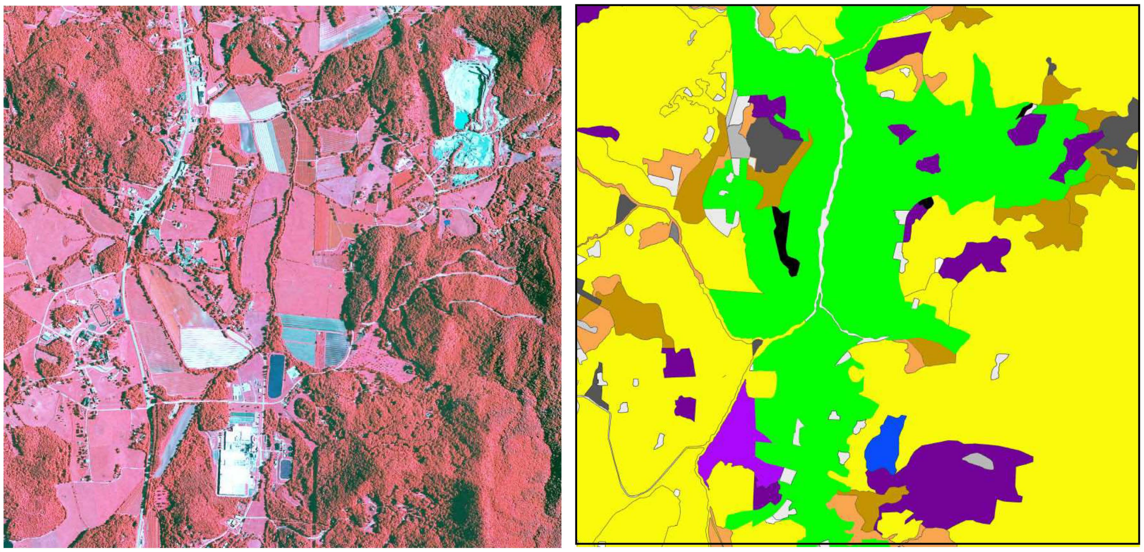


Figure 1-7: Images taken from aircraft and satellites (left) provide a rich source of data, which may be interpreted and converted to information about the Earth's surface (right).

GIS in Action

Spatial data organization, analyses, and delivery are widely applied to improve life. Here we describe examples that demonstrate how GIS are in use.

Marvin Matsumoto was saved with the help of GIS. The 60-year-old hiker became lost in Joshua Tree National Park, a 300,000-hectare desert landscape famous for its distinct and rugged terrain. Between six and eight hikers become lost there in a typical year, sometimes fatally so. Because of the danger of hypothermia, dehydration, and death, the U.S. National Park Service (NPS) organizes search and rescue operations that include foot patrols, horseback, vehicle, and helicopter searches (Figure 1-8).

The search and rescue operation for Mr. Matsumoto was organized and guided using GIS. Search and rescue teams carried field positioning devices that recorded team location and progress. Position data were downloaded from the field devices to a field GIS center, and frequently updated maps were produced. On-site incident managers used these maps to evaluate areas that had been searched, and to plan subsequent efforts in real time. Accurate maps showed exactly what portions of the park had been searched and by what method. Appropriate teams were tasked to unvisited areas. Ground crews could be assigned to areas that had been searched by helicopters, but contained vegetation or terrain that limited visibility from above. Marvin was found on the fifth day, alive but dehydrated and with an injured skull and back from a fall. The search team was able to radio its precise location to a rescue helicopter. Another day in the field and Marvin likely would have died, a day saved by the effective use of GIS.

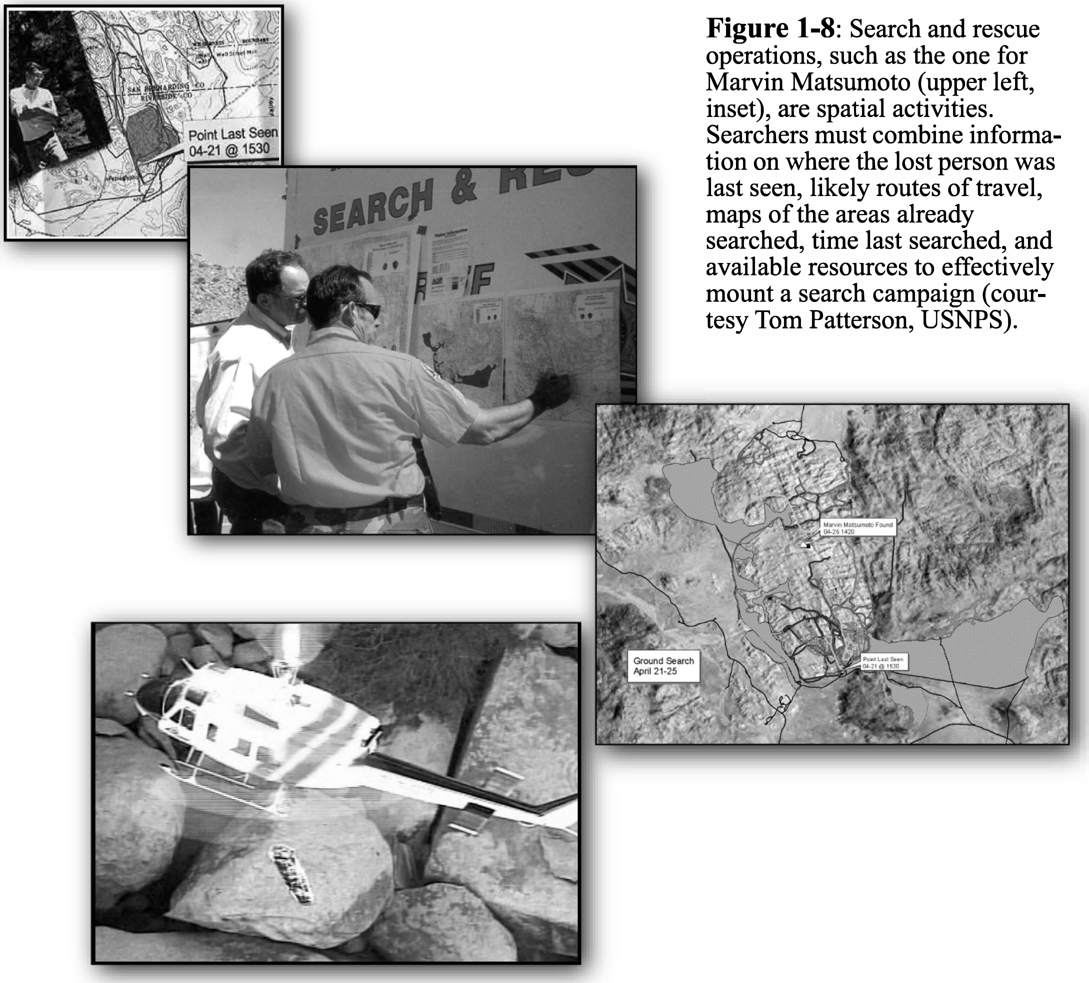


Figure 1-8: Search and rescue operations, such as the one for Marvin Matsumoto (upper left, inset), are spatial activities. Searchers must combine information on where the lost person was last seen, likely routes of travel, maps of the areas already searched, time last searched, and available resources to effectively mount a search campaign (courtesy Tom Patterson, USNPS).

GIS are also widely used in planning and environmental protection. Oneida County is located in northern Wisconsin, a forested area characterized by exceptional scenic beauty. The county is in a region with among the highest concentrations of freshwater lakes in the world, a region that is also undergoing a rapid expansion in the permanent and seasonal human populations. Retirees, urban exiles, and vacationers are increasingly drawn to the scenic and recreational amenities available in Oneida County. Permanent county population grew by nearly 30% from 1990 to 2010, and the seasonal influx almost doubles the total county population each summer.

Population growth has caused a boom in construction and threatened the lakes that draw people to the county. A growing number of building permits are for nearshore houses, hotels, or businesses. Seepage from septic systems, runoff from fertilized lawns, or erosion and sediment from construction all decrease lake water quality. Increases in lake nutrients or sediment may lead to turbid waters, reducing the beauty and value of the lakes and nearby properties.

In response to this problem, Oneida County, the Sea Grant Institute of the University of Wisconsin, and the Land Information and Computer Graphics Facility of the University of Wisconsin have developed a Shoreland Management GIS Project. This project helps protect valuable nearshore and lake resources, and provides an example of how GIS tools are used for water resource management (Figure 1-9).

Oneida County has revised zoning and other ordinances to protect shoreline and lake quality, and to ensure compliance without undue burden on landowners. The county uses GIS technology in the maintenance of property records. Property records include information on the owner, tax value, and any special zoning considerations. The county uses these digital records when creating parcel maps; processing sale, subdivision, or other parcel transactions; and integrating new data such as aerial or boat-

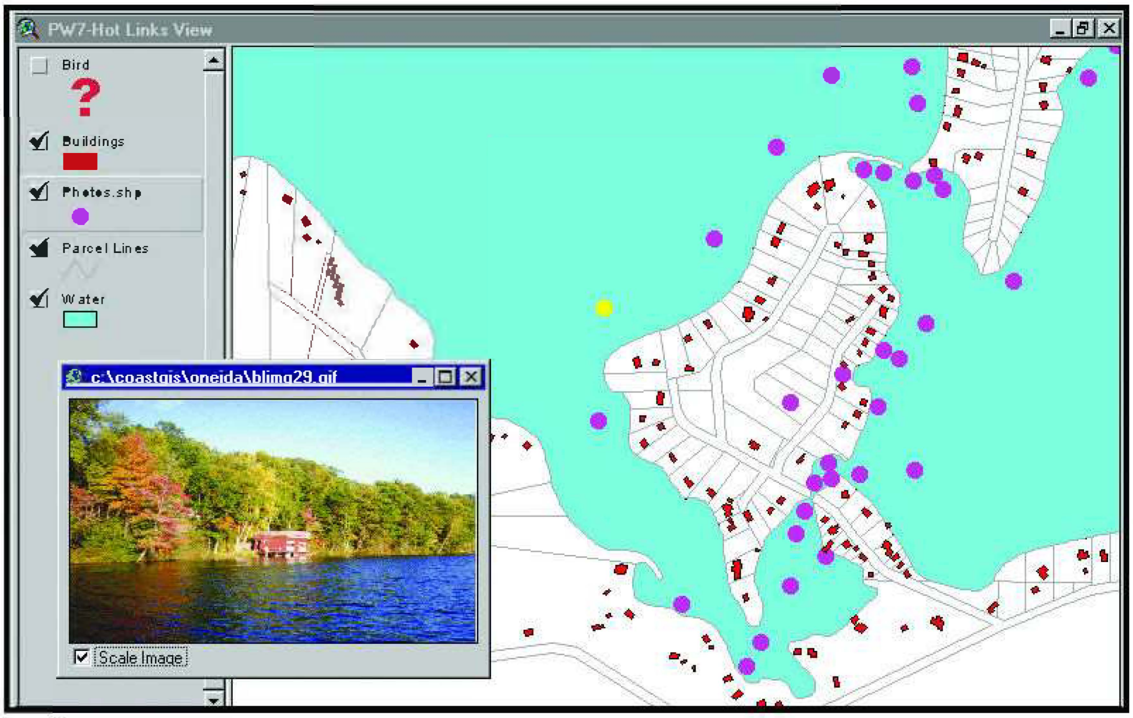


Figure 1-9: Parcel information entered in a GIS may substantially improve government services. Here, images of the shoreline taken from lake vantage points are combined with digital maps of the shoreline, buildings, and parcel boundaries. The image in the lower left was obtained from the location shown as a light dot near the center of the figure (courtesy Wisconsin Sea Grant Institute and LICGF).

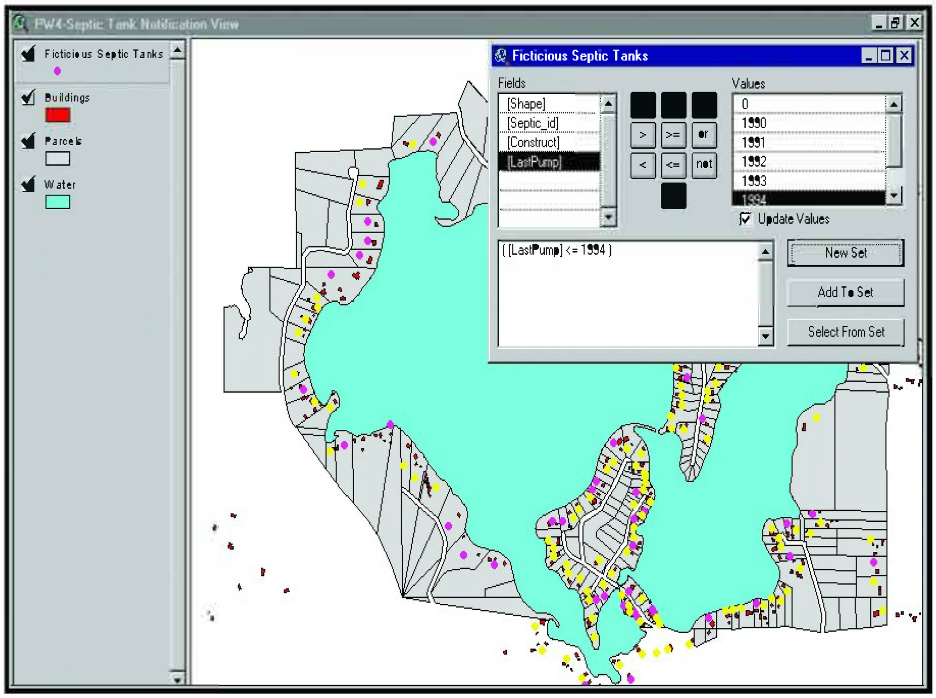


Figure 1-10: GIS may be used to streamline government function. Here, septic systems not compliant with pollution prevention ordinances are identified by light circles (courtesy Wisconsin Sea Grant Institute and LICGF).

based images to help detect property changes and zoning violations.

GIS may also be used to administer shoreline zoning ordinances, or to notify landowners of routine tasks, such as septic system maintenance. Northern lakes are particularly susceptible to nutrient pollution from nearshore septic systems (Figure 1-10). Timely maintenance of each septic system must be verified. The GIS can automatically identify owners out of compliance and generate an appropriate notification.

GIS has helped the U.S. Fish and Wildlife Service manage the recovery of the Gray Wolf (*Canis lupus*) in the lower 48 states of the United States. Wolves were hunted to a remnant population in northern Minnesota. Given protection in 1974, the population has rebounded to nearly 6,000 wolves that are spread across at least 11 states. GIS helped in many phases of the recovery, including identifying suitable habitat, monitoring pack location through time, mapping prey abundance and areas of high potential conflict with humans due to land use (e.g., ranching), assessing the impacts of range recovery on

Figure 1-11: A gray wolf, one of a few successfully recovered endangered species, restored with the help of GIS (courtesy Spinus Art Photos).

Figure 1-12: Wolf recovery involved tranquilizing and fitting wolves with tracking collars. These provide detailed location and movement data, and a better understanding of wolf habitat requirements (courtesy NPS).

other resources (deer and other game), and natural limits to range expansion

Relatively new spatial data capture technologies are used to help in wolf recovery. Animals are tranquilized, fitted with satellite tracking collars, and released (Figure 1-13). These collars may create an hour by hour record of wolf location, giving precise information on habitat occupancy, movement rates, hunting vs. resting time, optimal denning sites, and dispersal. More data are provided in a few weeks by these satellite tracking collars than were possible with a decade of collection using the older, radio-based technologies they replaced.

Scientists at the Voyageurs Wolf Project have been tracking wolves to better understand their behavior (Figure 1-13). Part of wolf recovery and de-listing may include hunting and trapping seasons in some areas. Harvest isn't allowed in U.S. National Parks, but may be on adjacent lands, e.g., State and National Forests. Removing pack members may affect a pack's ability to group hunt, reproduce, or defend their territory. Wolves may respond to hunting pressure by moving further into parks, in turn displacing adjacent packs. Analysis of pack location and movements during trial hunting and trapping may help guide a sustainable recovery.

GIS are widely used to improve public health. Air pollution is a major cause of sickness and death, primarily from nitrogen and sulfur dioxides, carbon monoxide, ozone, and small particles from oil, gas, coal, and

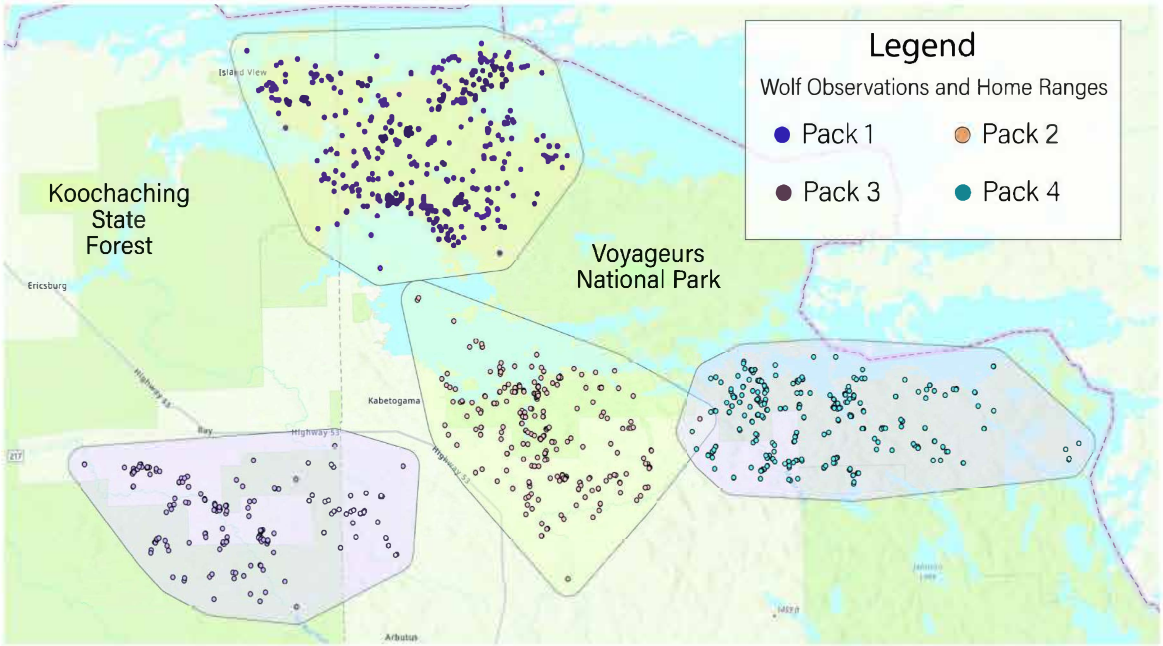


Figure 1-13: Spatial data, such as the recorded pack locations and home ranges (circles and polygons, respectively), may be combined to help understand how best to manage the Gray Wolf (courtesy Voyageurs Wolf Project).

Figure 1-14: Air pollution from power plants and vehicles is still a significant health hazard (courtesy M. Riya and the State of California).

wood combustion. Primary sources are power generation, factories, and transporation (Figure 1-14). Small particles lodge in the lungs , causing inflammation and reducing lung function (Figure 1-15). Alveolar macrophages attempt to isolate this material, but air pollution levels commonly exceed the lung's capacity for self-cleaning. Damaging particle concentrations are typically higher in urban areas, or near traffic, power plants, and other pollution sources. GIS helps map concentrations, identify sources, and plan improvements. Air pollution shaves 10 years off of the life span of about 200,000 people in the United States each year, and is responsible for the death of 7 million people worldwide each year. It also causes increased sickness, hospitalization, and medical costs that annually reach into the billions of dollars. A reduction in air pollution has been shown to significantly reduce hospitalization, childhood asthma, and to increase life expectancy.

Reducing sickness and death requires identifying areas of high exposure, particularly for vulnerable populations. Effective management requires an estimate of how much a decrease in pollution will increase health. Scientists have focused on these questions over the past decades, and can map exposures both over broader areas and at increasing level of spatial detail.

Air pollution may be mapped from satellites, as the chemicals and particles change the optical properties of air (Figure 1-16, top). A number of satellite instruments, culminating in the Ozone Mapping and Profiling Suite (OMPS), have been launched over the past 30 years to record air quality. Painstaking engineering, testing, and comparison to ground and airborne measurements have verified instrument accuracies. This has led

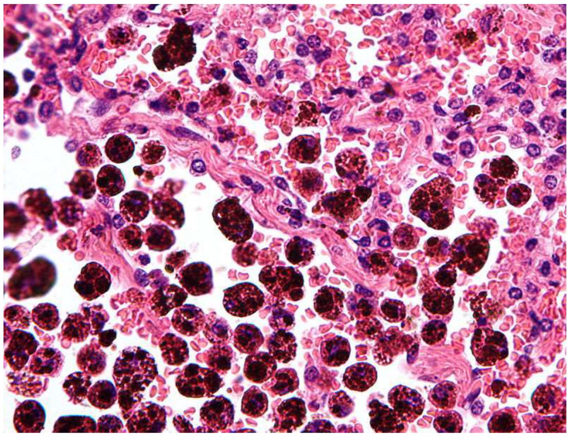
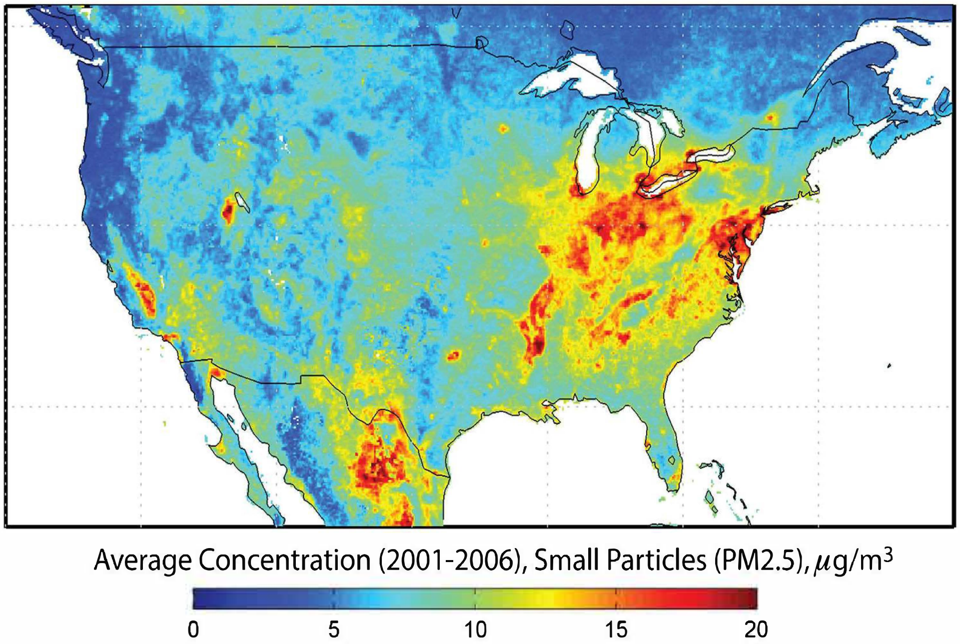


Figure 1-15:: Small air pollution particles (dark spots, above) lodge in lungs and cause life-long damage (courtesy Nephron).



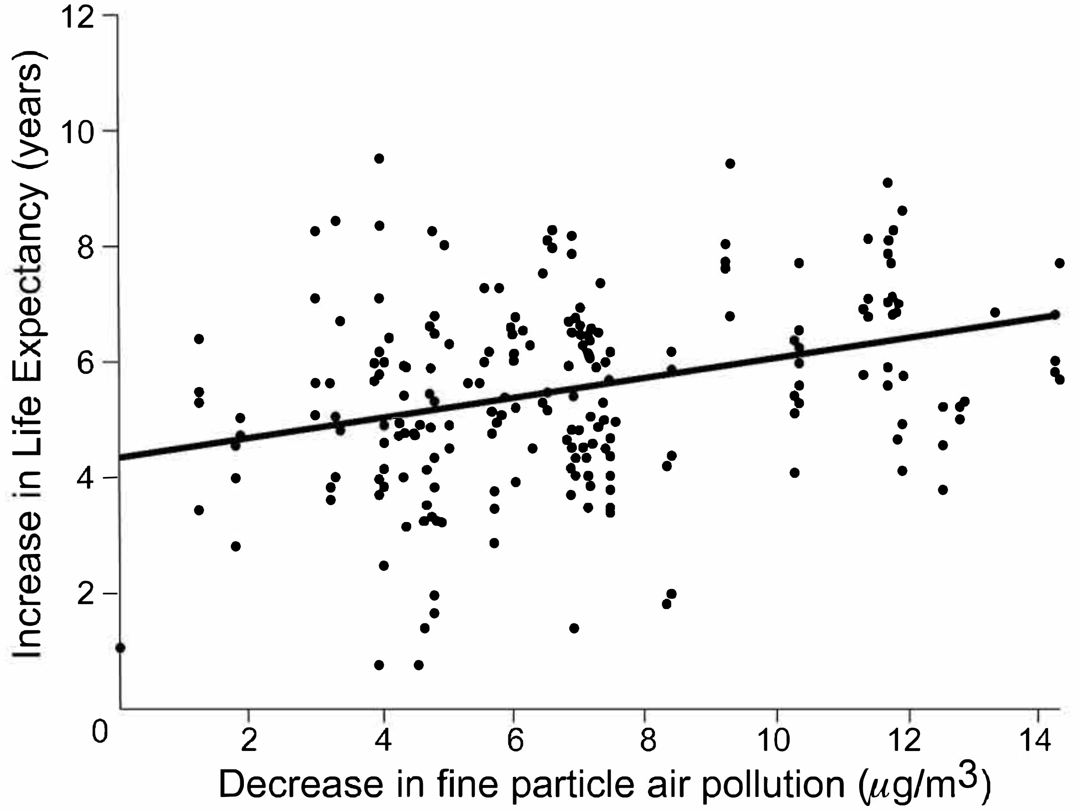


Figure 1-16: Scientists at NASA have developed methods to map air pollution across continents on a daily basis, which may be averaged to estimate chronic exposure (top). These spatial data may be combined with studies on human response to air pollution (bottom) and the location of vulnerable populations to improve public health and reduce medical costs.

to a long-term record of pollutant concentrations, and improved understanding of the sources and dynamics of pollutants across regional through global geographies. These data allow measurement of peak and chronic exposure to pollutants for different populations. They show persistent areas of high exposure (Figure 1-16), some concentrated in cities, largely due to automobile traffic, and others over large areas, e.g., the Midwest, due to large coal-fired power plants and industrial sources. Some areas are particularly prone to high concentrations due to surrounding highlands, e.g., the Central Valley of California or Salt Lake City, Utah.

Work by health scientists has identified the specific impacts of air pollution by analyzing response in target populations. Increased rates of asthma, lung damage, and death observed in smaller studies or individual cities can be expanded to broader areas through the combination of data in GIS. For example, combining health and population data with satellite exposure records has helped estimate the increase in life expectancy with a decrease in air pollution. Legislation passed in the 1970s resulted in a measurable improvement in air quality across the United States. Progress has been variable across the country, with some populations seeing larger reductions. Scientists measured the decrease in death rates in comparable populations, and estimated an average 2-year increase in life span for each 10 $\mu g\text{-}m^{-3}$ reduction in exposure (Figure 1-16, bottom).

Additional work has focused on air pollution at greater geographic detail, in part to better quantify and manage individual exposure and risk. Dr. Julian Marshall and collaborators at the University of Minnesota have developed systems to sample pollutant concentrations at very fine spatial intervals, towing an air sampling system behind a bicycle through a range of traffic densities, road types, and neighborhoods (Figure 1-17). Satellite positioning was synchronized with video and air samples, and these com-

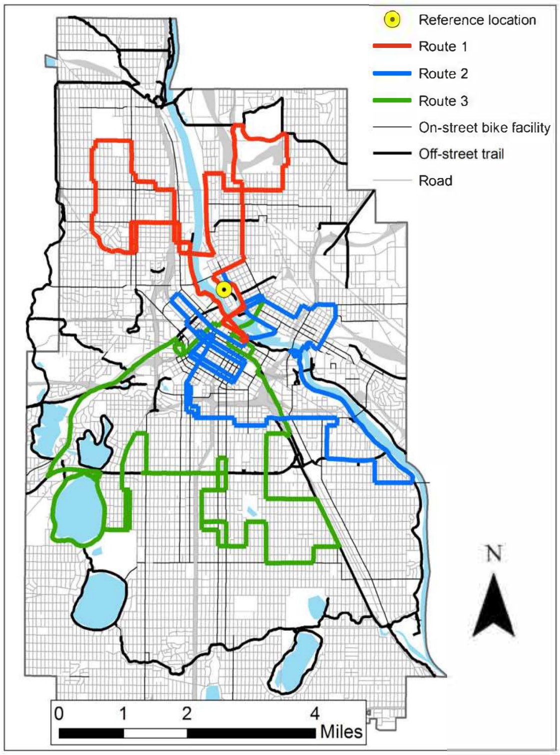


Figure 1-17: Towable samplers help measure air pollution for individual streets, at various traffic densities and types (courtesy J. Marshall).

bined with spatial data on road networks, population density, land use, and other factors. Statistical models were then developed. These allow detailed estimates of pollutant concentrations, even down to the individual street (Figure 1-18). Such estimates may in turn help reduce air pollution, plan bicycle or pedestrian corridors, separate the pollutant loadings due to cars vs. trucks, buses or other large vehicles, and manage traffic or infrastructure to reduce human exposure.

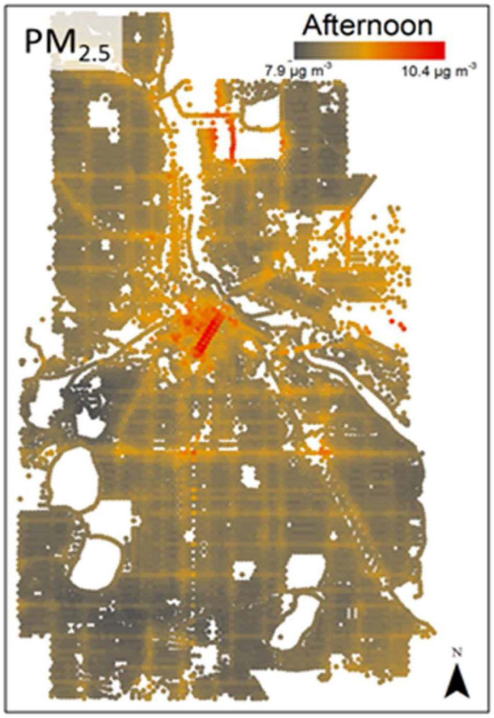


Figure 1-18: Fine-detailed spatial estimates of particulate air pollutants (courtesy J. Marshall).

GIS Components

A GIS is composed of hardware, software, data, humans, and a set of organizational protocols. These components must be well integrated for effective use of GIS, and the development and integration of these components is an iterative, ongoing process. The selection and purchase of hardware and software is often the easiest and quickest step in the development of a GIS. Data collection and organization, personnel development, and the establishment of protocols for GIS use are often more difficult and time-consuming endeavors.

Hardware for GIS

A fast computer, large data storage capacities, and a high-quality, large display form the hardware foundation of most GIS (Figure 1-19). A fast computer is required because spatial analyses are often applied over large areas and/or at high spatial resolutions. Calculations often have to be repeated over tens of millions of times, corresponding to each space we are analyzing in our geographical analysis. Even simple operations may take substantial time on general-purpose computers when run over large areas, and complex operations can be unbearably long-running. While advances in computing technology during the past decades have substantially reduced the time required for most spatial analyses, computation times are still unacceptably long for a few applications.

While most computers and other hardware used in GIS are general-purpose and adaptable for a wide range of tasks, there are also specialized hardware components that are specifically designed for use with spatial data. GIS require large volumes of data that must be entered to define the shape and location of geographic features, such as roads, rivers, and parcels. Specialized equipment, described in Chapters 4 and 5, has been developed to aid in these data entry tasks.

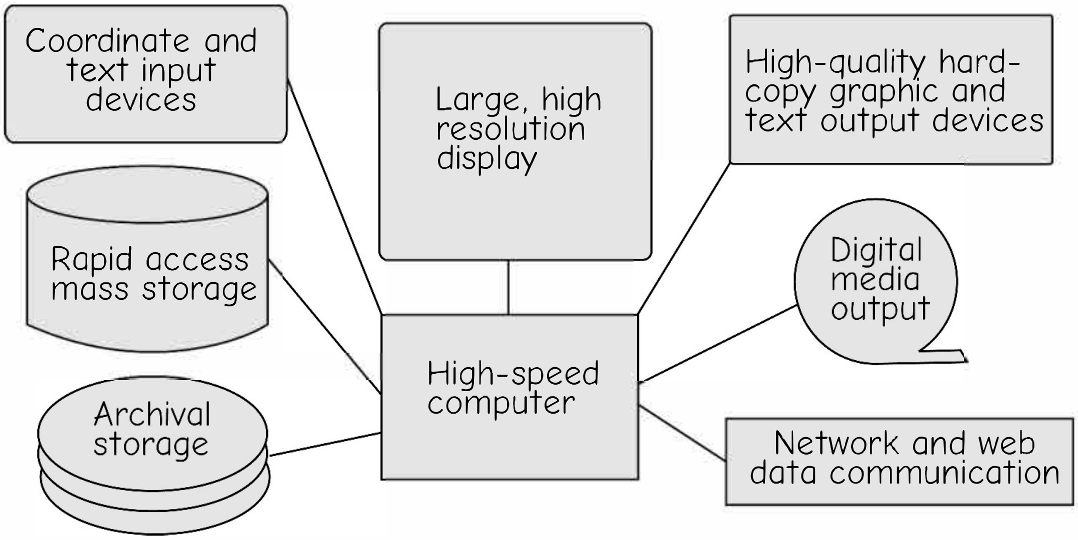


Figure 1-19: GIS are typically used with a number of general-purpose and specialized hardware components.

GIS Software

GIS software provides the tools to manage, analyze, and effectively display and disseminate spatial information (Figure 1-20). GIS by necessity involves the collection and manipulation of coordinates. We also must collect qualitative or quantitative information on the nonspatial attributes of geographic features. We need tools to view and edit these data, manipulate them to generate and extract the information we require, and produce the materials to communicate the information we have developed. GIS software provides the specific tools for some or all of these tasks.

There are many public domain and commercially available GIS software packages, and many of these packages originated at academic or government-funded research laboratories. The Environmental Systems Research Institute (ESRI) line of products, including ArcGIS, is a good example. Much of the foundation for early ESRI software was developed during the 1960s and 1970s at Harvard University in the Laboratory of Computer Graphics and Spatial Analysis. Alumni from Harvard included these in commercial products, and have developed additional methods and integrated new academic research in the five decades since.

Open Geospatial Consortium

We will briefly cover the most common GIS software, but first wish to introduce the Open Geospatial Consortium (OGC). Their efforts have eased sharing across various GIS softwares and computer operating systems. Standards for data formats, documentation, program interactions, and transmission have been developed and published (www.opengeospatial.org), and lists of standards-compliant software compiled. While some data structures remain opaque or proprietary, most have become open, and common standards ease community adoption, reduce barriers to switching among softwares, or adopting multiple geospatial processing packages. Compliance with the standards is a plus from a user's perspective, so a quick review of the OGC-compliant list is recommended when selecting a software platform.

Data entry
- manual coordinate capture
- attribute capture
- digital coordinate capture
- data import

Editing
- manual point, line and area feature editing
- manual attribute editing
- automated error detection and editing

Data management
- copy, subset, merge data
- versioning
- data registration and projection
- summarization, data reduction
- documentation
- compression
- indexing

Analysis
- spatial query
- attribute query
- interpolation
- connectivity
- proximity and adjacency
- buffering
- terrain analyses
- boundary dissolve
- spatial data overlay
- moving window analyses
- map algebra

Output
- map design and layout
- hardcopy map printing
- digital graphic production
- export format generation
- metadata output
- digital map serving

Figure 1-20: Functions commonly provided by GIS software.

Our software descriptions include the most widely used software packages, but are not all-inclusive. There are many additional software tools and packages available, particularly for specialized tasks or subject areas.

ArcGIS

ArcGIS, in its various online, desktop, and server versions, comprises the most popular GIS software suite at the time of this writing. ESRI, the developer of ArcGIS, has a worldwide presence. ESRI has been producing GIS software since the early 1980s, and ArcGIS is its most recent and well-developed integrated GIS package. In addition to software, ESRI also provides substantial training, support, and fee-consultancy services at regional and international offices.

ArcGIS is designed to provide a large set of geoprocessing procedures, from data entry through analysis to most forms of data output. As such, ArcGIS is a large, complex, sophisticated product. It supports multiple data formats, many data types and structures, and literally thousands of possible operations that may be applied to spatial data. It is not surprising that substantial training is required to master the full capabilities of ArcGIS.

ArcGIS provides wide flexibility in how we conceptualize and model geographic features. Geographers and other GIS-related scientists have conceived of many ways to think about, structure, and store information about spatial objects. ArcGIS provides for the broadest available selection of these representations. For example, elevation data may be stored in at least four major formats, each with attendant advantages and disadvantages. There is equal flexibility in the methods for spatial data processing. This broad array of choices, while responsible for the large investment in time required for mastery of ArcGIS, provides concomitantly substantial analytical power.

QGIS

QGIS is an open-source software project, an initiative under the Open Source Geospatial Foundation. The software is a collaborative effort by a community of developers and users. QGIS is free, stable, changes smoothly through time, with the source code available so that it can be extended as needed for specific tasks. It provides a graphical user interface, supports a wide variety of data types and formats, and runs on Unix, MacOSX, and Microsoft Windows operating systems. As with most open-source software, the original offering had limited capabilities. With an average of approximately two updates a year since 2002, QGIS provides a large number of basic GIS display and analysis functions. An interface has been developed with GRASS, another open-source GIS with complementary analytical functions, but that lacks as straightforward a graphical user interface.

GeoMedia

GeoMedia and related products are the popular GIS suite from Hexagon Geospatial. GeoMedia offers a complete set of data entry, analysis, and output tools. A comprehensive set of editing tools may be purchased, including those for automated data entry and error detection, data development, data fusion, complex analyses, and sophisticated data display and map composition.

GeoMedia is particularly adept at integrating data from divergent sources, formats, and platforms. Intergraph appears to have dedicated substantial effort toward the OpenGIS initiative, a set of standards to facilitate cross-platform and cross-software data sharing. Data in any of the common commercial databases may be integrated with spatial data from many formats. Image, coordinate, and text data may be combined.

GeoMedia also provides a comprehensive set of tools for GIS analyses. Complex spatial analyses may be performed, including queries, for example, to find features in the database that match a set of conditions,

and spatial analyses such as proximity or overlap between features. World Wide Web and mobile phone applications are well supported.

MapInfo

MapInfo is a comprehensive set of GIS products developed by the MapInfo Corporation, but now a part of Pitney Bowes. MapInfo products are used in a broad array of endeavors, although use seems to be concentrated in many business and municipal applications. This may be due to the ease with which MapInfo components are incorporated into other applications. Data analysis and display components are supported through a range of higher language functions, allowing them to be easily embedded in other programs. In addition, MapInfo provides a flexible, stand-alone GIS product that may be used to solve many spatial analysis problems.

Specific products have been designed for the integration of mapping into various classes of applications. For example, MapInfo products have been developed for embedding maps and spatial data into wireless handheld devices such as telephones, data loggers, or other portable devices. Products have been developed to support internet mapping applications, and serve spatial data in World Wide Web-based environments. Extensions to specific database products such as Oracle are provided.

Idrisi

Idrisi is a GIS system developed by the Graduate School of Geography of Clark University, in Massachusetts. Idrisi differs from the previously discussed GIS software packages in that it provides both image processing and GIS functions. Image data are useful as a source of information in GIS. There are many specialized software packages designed specifically to focus on image data collection, manipulation, and output. Idrisi offers much of this functionality while also providing a large suite of spatial data analysis and display functions.

Idrisi has adopted a number of very simple data structures, a characteristic that makes the software easy to modify. Some of these structures, while slow and more space-demanding, are easy to understand and manipulate for the beginning programmer. The space and speed limitations have become less relevant with improved computers. File formats are well documented and data easy to access. The developers of Idrisi have expressly encouraged researchers, students, and users to create new functions for Idrisi. Idrisi is an ideal package for teaching students both to use GIS and to develop their own spatial analysis functions.

A suite of tools for earth system modeling has been developed on the Idrisi platform, and combined in the Tereset software system. Functions include land change modeling, habitat and biodiversity modeling, and climate change adaptation.

Manifold

Manifold is a relatively inexpensive GIS package with a surprising number of capabilities. Manifold combines GIS and some remote sensing capabilities. Basic spatial data entry and editing support are provided, as well as projections, basic vector and raster analysis, image display and editing, and output. The program is extensible through a series of software modules. Modules are available for surface analysis, business applications, internet map development and serving, database support, and advanced analyses.

Manifold GIS has focused on rapid computations for large spatial databases, and in providing sophisticated image editing capabilities in a spatially referenced framework. Portions of images and maps may be cut and pasted into other maps while maintaining proper geographic alignment. Transparency, color-based selection, and other capabilities common to image editing programs are included in Manifold GIS.

AUTOCAD MAP 3D

AUTOCAD is the world's largest-selling computer drafting and design package. Produced by Autodesk, Inc. of San Rafael, California, AUTOCAD began as an engineering drawing and printing tool. A broad range of engineering disciplines are supported, including surveying and civil engineering. Surveyors have traditionally developed and maintained the coordinates for property boundaries, and these are among the most important and often-used spatial data. AUTOCAD MAP 3D adds substantial analytical capability to the already complete set of data input, coordinate manipulation, and data output tools provided by AUTOCAD.

GRASS

GRASS, the Geographic Resource Analysis Support System, is a free, open-source GIS that runs on many platforms. The system was originally developed by the U.S. Army Construction Engineering Research Laboratory (CERL), starting in the early 1980s, when much GIS software was limited in access and applications. CERL followed an open approach to development and distribution, leading to substantial contributions by a number of university and other government labs. Development was discontinued by the military, and taken up by an open-source "GRASS Development Team," a self-identified group of people donating their time to maintain and enhance GRASS. The software provides a broad array of raster and vector operations, and is used in both research and applications worldwide. Detailed information and the downloadable software are available at http://grass.itc.it/index.php.

MicroImages

MicroImages produces TNTmips, an integrated remote sensing, GIS, and CAD software package. MicroImages also produces and supports a range of other related products, including software to edit and view spatial data, software to create digital atlases, and software to publish and serve data on the internet.

TNTmips is notable both for its breadth of tools and the range of hardware platforms supported in a uniform manner. MicroImages recompiles a basic set of code for each platform so that the look, feel, and functionality is nearly identical irrespective of the hardware platform used. Image processing, spatial data analysis, and image, map, and data output are supported uniformly across this range.

TNTmips provides an impressive array of spatial data development and analysis tools. Common image processing tools are available, including support of a broad number of file formats, image registration and mosaics, reprojection, error removal, subsetting, combination, and image classification. Vector and raster analyses are supported, including multi-layer combination, viewshed, proximity, and network analyses. Extensive online documentation is available, and the software is supported by an international network of dealers.

ERDAS

ERDAS (Earth Resources Data Analysis System) – now owned and developed by Hexagon Geospatial, a division of Intergraph – began as an image processing system. The original purpose of the software was to enter and analyze satellite image data. ERDAS led a wave of commercial products for analyzing spatial data collected over large areas. Product development was spurred by the successful launch of the U.S. Landsat satellite in the 1970s. For the first time, digital images of the entire Earth surface were available to the public.

The ERDAS software evolved to include a comprehensive set of tools for cell-based data analysis. Image data are supplied in a cell-based format. The "checkerboard" format used for image data may also be used to store and manipulate other spatial data.

ERDAS and most other image processing packages provide data output formats that are compatible with most common GIS packages. Many image processing software systems are purchased explicitly to provide data for a GIS. The support of ESRI data formats is particularly thorough in ERDAS. ERDAS GIS components can be used to analyze these spatial data.

ENVI

ENVI is another GIS software package with origins in digital image processing. Particular emphasis has been placed on tools for developing and managing elevation data from satellites and airborne platforms, crop monitoring, and automated feature extraction. This last capability streamlines the identification of individual objects, such as buildings, trees, road segments, or water bodies. Recent updates have focused on tools for processing images from small, unmanned aerial drones.

Bentley Map

Bentley Systems has developed spatial analysis software for mobile device through enterprise levels, with a strong focus on flexible, integrated infrastructure design and development. Although its origins are as a computer-assisted drafting and design program, Bentley has evolved into a general set of tools, including field data collection, photogrammetry, sophisticated map composition, database management, analysis, and reporting.

Bentley products are particularly focused on the built environment, including road, building, utility, and other large construction design, planning, and management. Tools include a comprehensive suite for property records, including surveying parcel data management, terrain analysis and calculations for excavation and earthworks, rainfall runoff analysis and drainage design, street and utility layout, and 3D viewing of design alternatives. Bentley also supports industry-specific tools, including mining and power generation systems and networks.

SuperMap

SuperMap is a Hong Kong based company that provides a broad range of GIS software, including desktop, cloud-based, vector, raster, and 3D analysis. Subsystems and configurations have been developed and applied for land records and information management, facilities management, government economic and statistical services and support, municipalities, and emergency response management. It provides excellent support of Japanese, Korean, and Chinese languages, and has among the largest market shares in East Asia.

Spatial R, Python, and GDAL

Generic programing, processing, and statistical analysis tools may be combined to provide most GIS functions, and include newer analytical methods not available in common commercial packages. R is an open source software project with many spatial packages. These support a rich set of spatial operations, particularly for spatial estimation. Python is a general-purpose programming language with several available spatial libraries. Notable among them are Shapely, Geopandas, and pySAL, containing a large set of spatial functions. GDAL is a standard set of spatial input/output and data processing functions, which may interfaces with both R and Python. Together, these tools support sophisticated GIS analysis.

This review of spatial data software is incomplete. There are many other software tools available which provide unique, novel, or particularly clever combinations of geoprocessing functions. Whitebox GAT, Smallworld, ILWIS, MapWindow, PCI, and qvSIG are just a few additional software packages with spatial data capabilities. In addition, there are thousands of add-ons, special-purpose tools, or specific modules that complement these products.

GIS in Organizations

Although new users often focus on GIS hardware and software components, we must recognize that GIS exist in an institutional context. Effective use of GIS requires an organization to support various GIS activities. Most GIS also require trained people to use them, and a set of protocols guiding how the GIS will be used. The institutional context determines what spatial data are important, how these data will be collected and used, and ensures that the results of GIS analyses are properly interpreted and applied. GIS share a common characteristic of many powerful technologies. If not properly used, GIS may lead to a significant waste of resources, and may do more harm than good. The proper institutional resources are required for GIS to provide all its potential benefits.

GIS are often employed as decision support tools (Figure 1-21). Data are collected, entered, and organized into a spatial database, and analyses performed to help make specific decisions. The results of spatial analyses in a GIS often uncover the need for more data, and there are often several iterations through the collection, organization, analysis, output, and assessment steps before a final decision is reached. It is important to recognize the organizational structure within which the GIS will operate, and how GIS will be integrated into the decision-making processes of the organization.

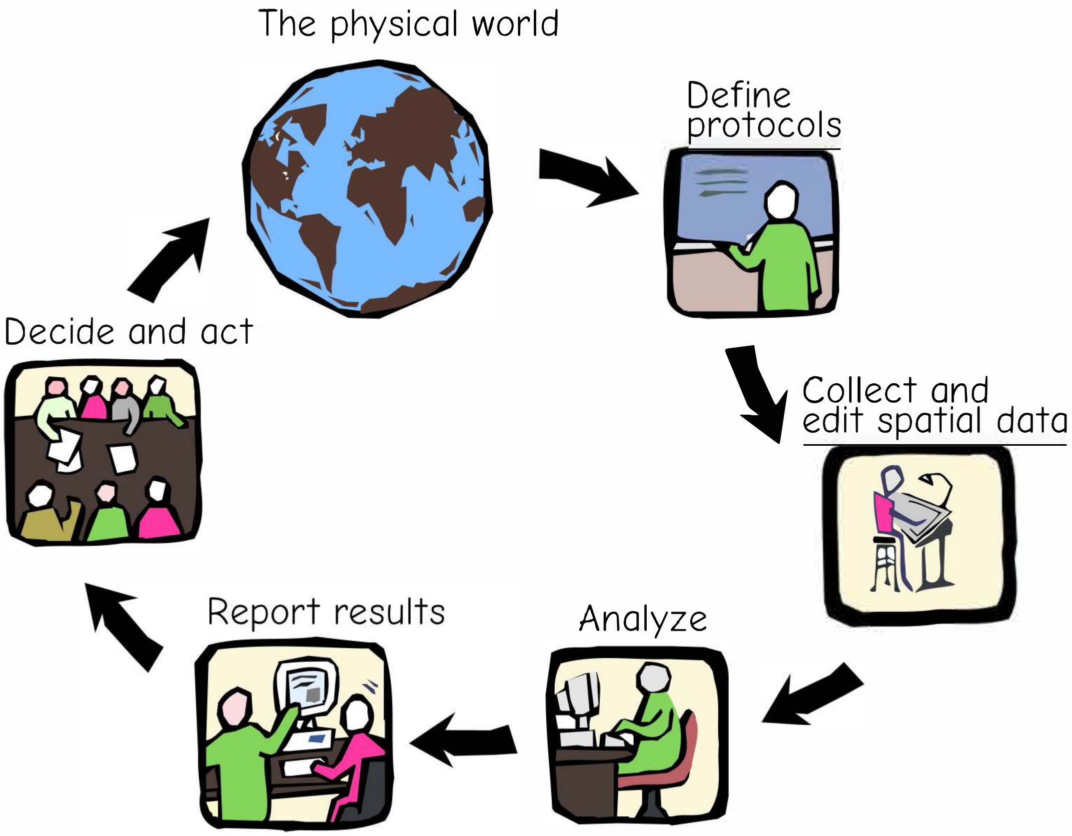


Figure 1-21: GIS exist in an institutional context. Effective use of GIS depends on a set of protocols and an integration into the data collection, analysis, decision, and action loop of an organization.

One first question is, “What problem(s) are we to solve with the GIS?” GIS add significant analytical power through the ability to measure distances and areas, identify vicinity, analyze networks, and through the overlay and combination of different information. Unfortunately, spatial data development is often expensive, and effective GIS use requires specialized knowledge or training, so there is often considerable expense in constructing and operating a GIS. Before spending this time and money, there must be a clear identification of the new questions that may be answered, or the process, product, or service that will be improved, made more efficient, or less expensive through the use of GIS. Once the ends are identified, an organization may determine the level of investment in GIS that is warranted.

Summary

GIS are computer-based systems that aid in the development and use of spatial data. There are many reasons we use GIS, but most are based on a societal push, our need to more effectively and efficiently use our resources. It also responds to a technological pull, our interest in applying new tools to previously insoluble problems. GIS as a technology is based on geographic information science, and is supported by the disciplines of geography, surveying, engineering, space science, computer science, cartography, statistics, and a number of others.

GIS are composed of both hardware and software components. Because of the large volumes of spatial data and the need to input coordinate values, GIS hardware often have large storage capacities, fast computing speed, and ability to capture coordinates. Software for GIS are unique in their ability to manipulate coordinates and associated attribute data. A number of software tools and packages are available to help us develop GIS.

While GIS are defined as tools for use with spatial data, we must stress the importance of the institutional context in which GIS fit. Because GIS are most often used as decision support tools, the effective use of GIS requires more than the purchase of hardware and software. Trained personnel and protocols for use are required if GIS are to be properly applied. GIS may then be incorporated in the question–collect–analyze–decide loop when solving problems.

The Structure of This Book

This book is designed to serve a semester-long, 15-week course in GIS at the university level. We seek to provide the relevant information to create a strong basic foundation on which to build an understanding of GIS. Because of the breadth and number of topics covered, students may be helped by knowledge of how this book is organized. Chapter 1 (this chapter) sets the stage, providing some motivation and a background for GIS. Chapter 2 describes basic data representations. It treats the main ways we use computers to represent perceptions of geography, common data structures, and how these structures are organized. Chapter 3 provides a basic description of coordinates and coordinate systems, how coordinates are defined and measured on the surface of the Earth, and conventions for converting these measurements to coordinates we use in a GIS.

Chapters 4 through 7 treat spatial data collection and entry. Data collection is often a substantial task and comprises one of the main activities of most GIS organizations. General data collection methods and equipment are described in Chapter 4. Chapter 5 describes Global Navigation Satellite Systems (GNSS), a common technology for coordinate data collection. Chapter 6 describes aerial and space-based images as a source of spatial data. Most historical and contemporary maps depend in some way on image data, and this chapter provides a background on how these data are collected and used to create spatial data. Chapter 7 provides a brief description of common digital data sources available in the United States, their formats, and uses.

Chapters 8 through 13 treat the analysis of spatial data. Chapter 8 focuses on attribute data, attribute tables, database design, and analyses using attribute data. Attributes are half our spatial data, and a clear understanding of how we structure and use them is key to effective spatial reasoning. Chapters 9, 10, 11, and 12 describe basic spatial analyses, including adjacency, inclusion, overlay, and data combination for the main data models used in GIS. They also describe more complex spatio-temporal models. Chapter 13 describes various methods for spatial prediction and interpolation. We typically find it impractical or inefficient to collect "wall-to-wall" spatial and attribute data. Spatial prediction allows us to extend our sampling and provide information for unsampled locations. Chapter 14 describes how we assess and document spatial data quality, while Chapter 15 provides some musings on current conditions and future trends.

We give preference to the International System of Units (SI) throughout this book. The SI system is adopted by most of the world, and is used to specify distances and locations in the most common global coordinate systems and by most spatial data collection devices. However, some English units are culturally embedded, for example, the survey foot, or 640 acres to a Public Land Survey Section, and so these are not converted. Because a large portion of the target audience for this book is in the United States, English units of measure often supplement SI units.

Suggested Reading

Ballas, D., Clarke, G., Franklin, R.S., Newing, A. (2018). *GIS and the Social Sciences: Theory and Applications.* London: Routledge.

Convis, C.L, Jr. (Ed.). (2001). *Conservation Geography: Case Studies in GIS, Computer Mapping, and Activism*. Redlands: ESRI Press.

Day, B., Bruner, J., Moser, A. (2017). *Geospatial Data and Analysis*. Sebastopol CA: O'Reilly.

Dent, B., Torguson, J.S., Hodler, T.W. (2009). *Cartography, Thematic Map Design*, 6th Edition. New York: McGraw Hill.

Dodge, M., McDerby, M., Turner, M. (2008). *Geographic Visualization: Concepts, Tools, and Applications*. Hoboken: Wiley.

Fotheringham, S., Rogerson, P.A. (2009). *The SAGE Handbook of Spatial Analysis*. London: SAGE.

Greene, R.P., Pick, J.B. (2012). *Exploring the Urban Community: A GIS Approach*. Upper Saddle River: Prentice Hall.

Haklay, M. (2010). *Interacting with Geospatial Technologies*. New York: Wiley.

Hankey, S., Marshall, J.D. (2016). Land use regression models of On-road Particulate Air Pollution (Particulate Number, Black Carbon, PM2.5, Particle Size) Using Mobile Monitoring. *Environmental Science and Technology,* 45:9194-9202.

Johnson, S. (2006). *The Ghost Map: the Story of London's Most Terrifying Epidemic, and How It Changed Science, Cities, and the World*. New York: Riverhead Books.

Kemp, K.K., (Ed.). (2008). *Encyclopedia of Geographic Information Science*. Los Angeles: SAGE.

Kouyoumjian, V. (2011). *GIS in the Cloud: The New Age of Cloud Computing and Geographic Information Systems*. Redlands: ESRI Press.

Kresse, W. Danko, D. M. (Ed.). (2012). *Handbook of Geographic Information*. Dordrecht: Springer.

Lawrence, P.L. (Ed.). (2013). *Geospatial Tools for Urban Water Resources*. New York: Dordrecht/Springer.

McHarg, I. (1995). *Design with Nature*. New York: Wiley.

Millspaugh, J.J., Thompson, F.R. III. (2009). *Models for Planning Wildlife Conservation in Large Landscapes*. Amsterdam: Elsevier.

Mueller, T., Sassenrath, G.F. (2015). *GIS Applications in Agriculture, Volume 4: Conservation Planning.* Boca Raton: CRC Press.

National Research Council of the National Academies (2006). *Beyond Mapping: Meeting National Needs through Enhanced Geographic Information Science*. Washington D.C.: National Academies Press.

Peterson, G.N. (2014). *GIS Cartography: A Guide to Effective Map Design*. Boca Raton: CRC Press.

Petrasova, A., Harmon, B., Petras, V., Mitasova, H. (2015). *Tangible Modeling with Open Sources GIS*. New York: Springer.

Shelito, B.A. (2012). *Introduction to Geospatial Technologies*. New York: W.H. Freeman.

Singleton, A.D., Spielman, S.E., Folch, D.C. (2018). *Urban Analytics*. Los Angeles: Sage.

de Smith, M.G., Goodchild, M.F., Longley, P.A. (2007). *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques, and Software Tools.* Leicester: Winchelsea Press.

Theobald, D. M. (2003). *GIS Concepts and ArcGIS Methods*. Fort Collins: Conservation Planning Technologies.

Tillman Lyle, J. (1999). *Design for Human Ecosystems: Landscape, Land Use, and Natural Resources*. Washington: Island Press.

Wegmann, M., Leutner, B., Dech, S. (2016). *Remote Sensing and GIS for Ecologists: Using Open Source Software*. Exeter: Pelagic Publishing.

Wise, S., Craglia, M. (Eds.). (2008). *GIS and Evidence-based Policy Making.* Boca Raton: CRC Press.

Exercises

1.1 - Why are we more interested in spatial data today than 100 years ago?

1.2 - You have probably collected, analyzed, or communicated spatial data in one way or another during the past month. Describe each of these steps for a specific application you have used or observed.

1.3 - How are GIS hardware different from most other hardware?

1.4 - Describe the ways in which GIS software are different from other computer software.

1.5 - What are the limitations of using a GIS? Under what conditions might the technology hinder problem solving, rather than help?

1.6 - Are paper maps and paper data sheets a GIS? Why or why not?

2 Data Models

Introduction

Data in a GIS represent a simplified view of physical *entities* – the roads, mountains, accident locations, or other features we care about. Data include information on the spatial location and nonspatial properties of entities.

Each entity is represented by a *spatial object* in a GIS, defining an entity-object correspondence. Because every computer system has limits, we can't save the exact boundary or all characteristics of features. As illustrated in Figure 2-1, we may represent land cover by a set of polygons. The polygon boundaries may be defined by a connected set of points, e.g., at an average spacing of approximately every 3 meters. We may record data that define each land cover, perhaps vegetation type, ownership, and landuse. Edge details smaller than 3 m and unrecorded characteristics such as value are not included in this representation.

The spatial detail and essential characteristics are subjectively chosen by the data developer. The density of points required by a surveyor will be different than that for a land use planner. The essential characteristics of a forest would be different in the eyes of a logger than those of a hunter or hiker. No one representation is universally better than any other, and the GIS developer seeks to define objects that support the intended use of the data, at the desired level of detail and accuracy.

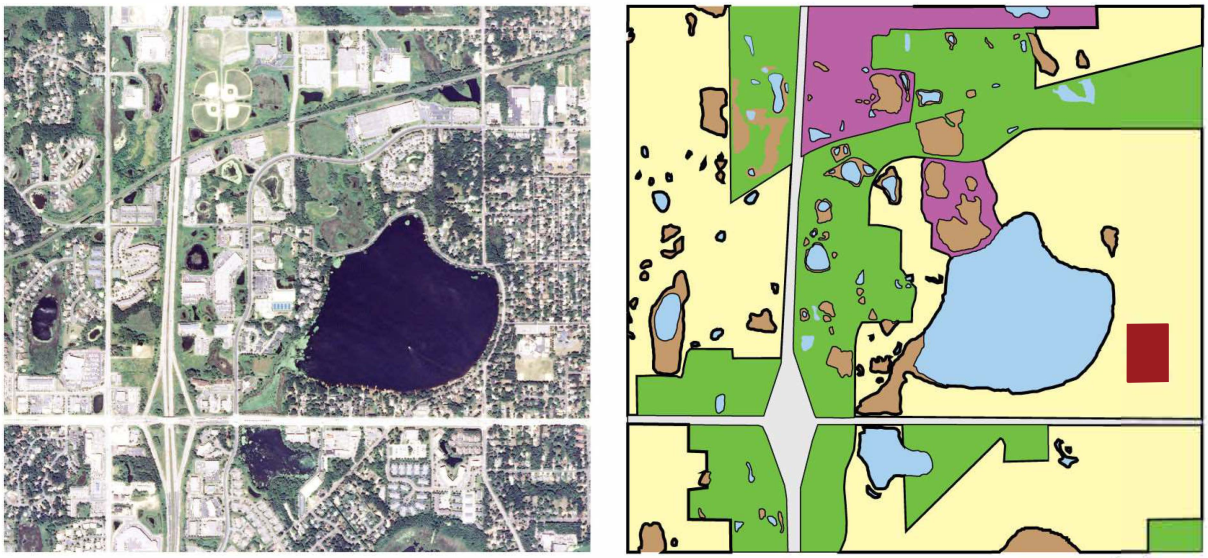


Figure 2-1: A physical entity is represented by a spatial object in a GIS. Here, lakes (dark areas in the photograph) and other land cover types are represented by polygons in the data layers on the right.

A *spatial data model* (Figure 2-2) may be defined as the objects in a spatial database plus the relationships among them. The term "model" is fraught with ambiguity because it is used in many disciplines to describe many things. Here, a spatial data model provides a formal means of representing and manipulating spatially referenced information. In Figure 2-1, our data model consists of two parts. The first is a set of polygons recording the edges of distinct land uses, and the second part (not shown in the figure) is a set of numbers, letters, or words associated with each polygon. The data model is the most recognizable level in our computer abstraction of the real world. Data structures (how we organize the information in the computer) and binary machine code (how we record it), are successively less recognizable but more computer-compatible forms of the spatial data (Figure 2-2).

Most GIS store our data as a set of layers (Figure 2-3). Each layer organizes the spatial and attribute data for a kind of cartographic object, and are often referred to as *thematic layers*. As an example, consider a GIS database that includes a soils data layer, a population data layer, an elevation data layer, and a roads data layer. The roads layer contains only roads data, including the location and properties of roads in the analysis area. Information on soils, political boundaries, and elevation are contained in their respective data layers. Through analyses we may combine data to create a new data layer; for example, we may identify areas that have high elevation and join this information with the soils data. This combination may create a new data layer with a composite soils-elevation variable.

Coordinates are used to define the spatial location and extent of geographic objects (Figure 2-4). A coordinate most often consists of a pair or triplet of numbers that specify location in relation to an origin. The coordinates quantify the distance from the

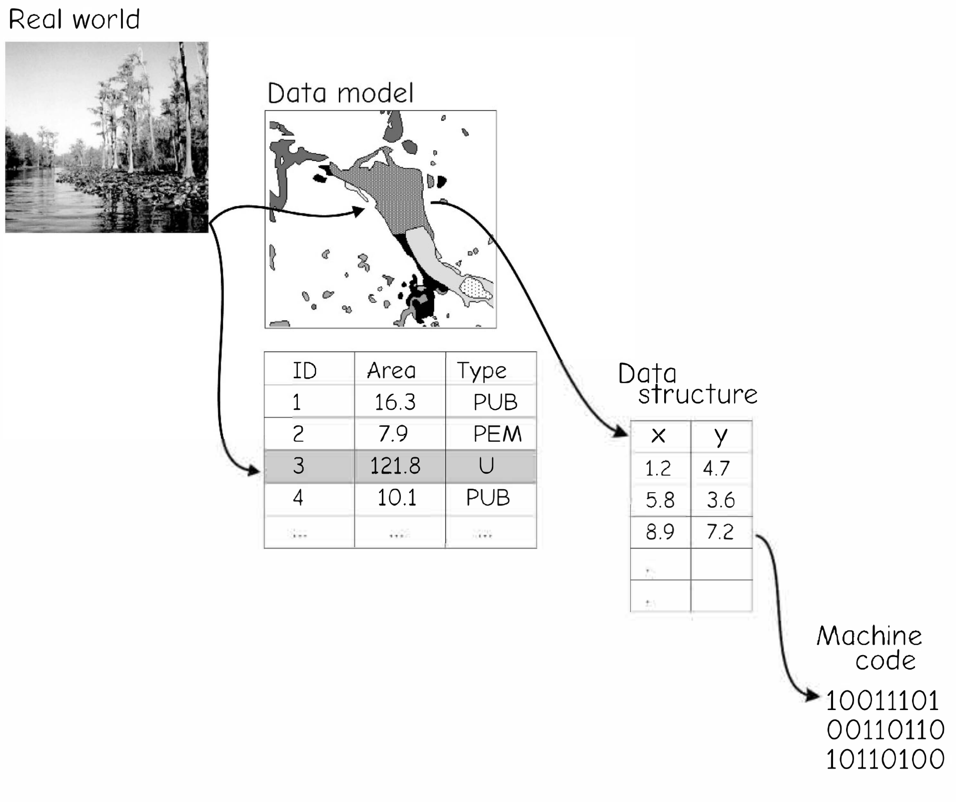


Figure 2-2: Levels of abstraction in the representation of spatial entities. The real world is represented in successively more machine-compatible but humanly obscure forms.

origin when measured along standard directions. Single or groups of coordinates are organized to represent the shapes and boundaries that define objects. Coordinates are usually based upon standardized map projections (discussed in Chapter 3). Each projection unambiguously defines the coordinate values for every point in an area.

Typically, attribute data complement the coordinate data for cartographic objects (Figure 2-4). These attribute data record the non-spatial components of an object, such as a name, color, pH, or cash value. Keys, labels, or other indexes are used so that the coordinate and attribute data may be viewed, related, and manipulated together.

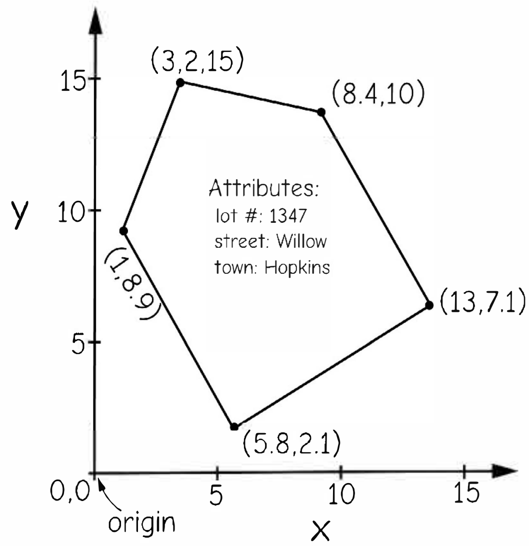


Figure 2-4: Coordinate and attribute data are used to represent entities.

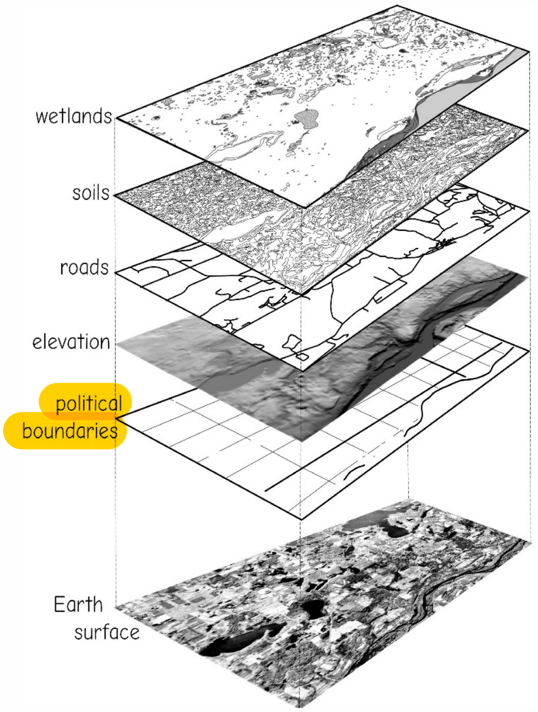


Figure 2-3: Spatial data are often stored as separate thematic layers, with objects grouped based on a set of properties, e.g., water, roads, or land cover, or some other agreed-upon set.

Coordinate Data

Coordinates define location in two- or three-dimensional space. Spatial data in a GIS most often use coordinate pairs, x and y, in a *Cartesian* coordinate system, named after Rene Descartes, the system's originator. These pairs define data on a flat, two-dimensional surface, and define the locations of features in our data layers.When working over large areas, we often require a three-dimensional representation. Coordinates in three dimensions are a bit more complicated because two alternate systems are common. Most adults are familiar with the concepts of latitude (ϕ), longitude (λ) and an elevation to define locations on the surface of the Earth. Spatial calculations are often easier in a three-dimensional Cartesian system starting near the Earth's center and using coordinate triplets X, Y, and Z. These alternate conventions for coordinate systems are described in turn in the following sections.

Planar Coordinate Systems

Planar, two-dimensional (2-D) Cartesian coordinate systems are the most common choice for GIS data storage and analysis. These systems define two *orthogonal* axes (right angle, or 90°), forming a plane (Figure 2-5). We specify a Y-axis, usually aligned at or close to a north-south direction, and an X-axis, usually aligned at or near an east-west direction. The Y-axis is often referred to as a *northing axis* and values increase upwards in a grid north direction. The X-axis is often referred to as an *easting axis* with values increasing to the right.

We must be careful when making measurements on our flat, 2-D data. When we display geographic data on a flat surface, we unavoidably distort relative locations, because the Earth's true surface is curved. Distance or area measurements are not the same on our imaginary flat surface as on the Earth's surface. We typically introduce small errors when we ignore the Earth's curvature, and we can keep errors below acceptably small values by limiting the area over which we use our flat 2-D model. As the mapped distance increases, the error increases to magnitudes we usually can't ignore. Specific methods for managing distortion in this curved to flat surface conversion are discussed in Chapter 3.

Coordinates on a Sphere

When we map over larger areas or when we need the highest precision and accuracy, we often use a three-dimensional, *spherical coordinate system*. Hipparchus, a Greek mathematician of the 2nd century B.C., was among the first to specify locations on the Earth using angular measurements on a sphere. A common spherical system uses two angles of rotation on a sphere with a fixed radius, R, to specify locations on Earth (Figure 2-6). The first angle of rotation, the longitude (λ), measures east-west distances around the polar, rotational axis of Earth. Zero is set for a line that passes near the Greenwich Observatory in England, and the distance angle is positive eastward and negative westward (Figure 2-6). The zero longitude, also known as the *Prime Meridian* or the *Greenwich Meridian*, was first specified through the Royal Greenwich Observatory in England, but measurement improvements, crustal movements, and changes in conventions now place zero longitude about 102 meters (335 feet) east of the Greenwich Observatory.

A second angle of rotation, measured along north-south lines that intersect the poles, is used to define a latitude (ϕ, Figure 2-6). Latitudes are specified as zero at the Equator, the line encircling the Earth that is

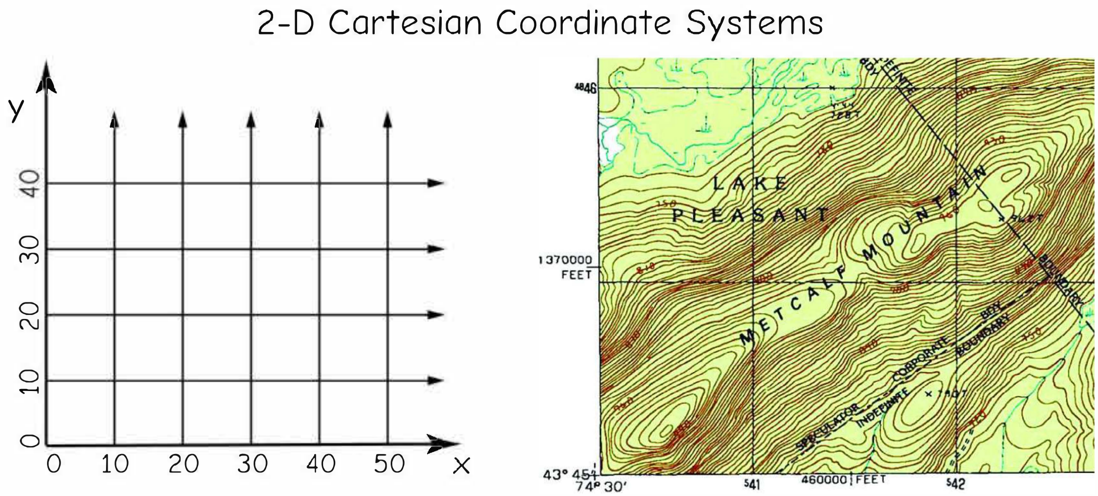


Figure 2-5: A 2-D coordinate system defines X and Y axes (left panel in figure above), and specify coordinate locations by these X-Y pairs. Coordinate values increase in rightward (X) and upward (Y) directions, and lines of constant X or Y values may be used to aid in location on maps (right, above).

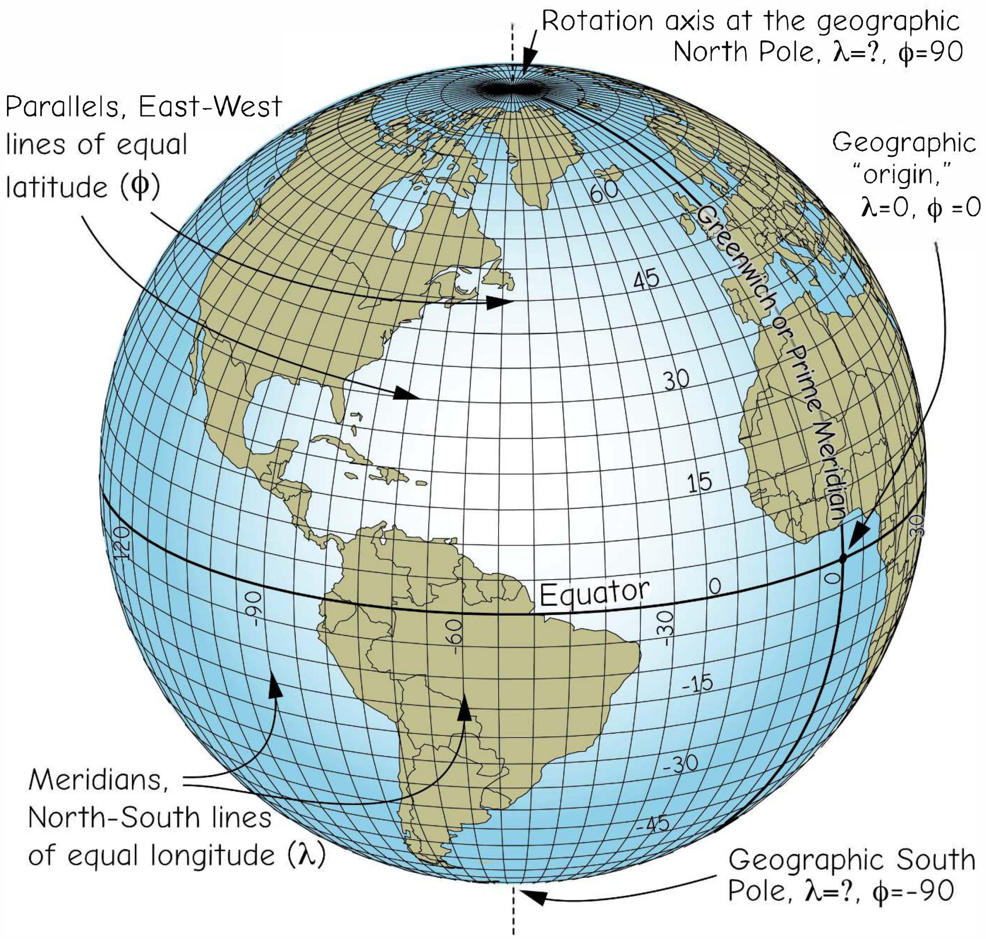


Figure 2-6: Conventions when referring to geographic latitudes and longitudes. Meridians are lines running north-south that have constant longitudes. Parallels are lines running east-west that have constant latitudes. Latitude is zero on the Equator. Longitude is zero on the Greenwich Meridian and undefined at the poles, because all longitudinal meridians intersect there (λ = ? in the figure).

always halfway between the North and South Poles. By convention, latitudes increase to maximum values of 90 degrees in the north and south, or, if a sign convention is used, from -90 at the South Pole to 90 at the North Pole. Lines of constant longitude are called meridians, and lines of constant latitude are called parallels (Figure 2-6). Because the meridians converge, geographic coordinates do not form a Cartesian system. A Cartesian system defines lines on a right-angle, planar grid. Geographic coordinates occur on a curved surface, and the longitudinal lines cross at the poles. This convergence means the distance spanned by a degree of longitude varies from approximately 111.3 kilometers at the Equator, to 0 kilometers at the poles. In contrast, the ground distance for a degree of latitude varies only slightly, from 110.6 kilometers at the Equator to 111.7 kilometers at the poles. The slight difference with latitude is due to a non-spherical Earth, something we'll describe a bit later.

Convergence causes distortion because a degree of latitude spans a greater distance near the poles than a degree of longitude. For example, "circles" with a fixed radius in geographic units, such as 5°, are not circles on the surface of the globe, with distortion greatest at the poles (Figure 2-7, left). They may appear as circles when the Earth's surface is "unrolled" and plotted on a flat map (Figure 2-7, right), but treating spherical coordinates (latitudes/longitudes) as Cartesian coordinates creates an inherently dis-

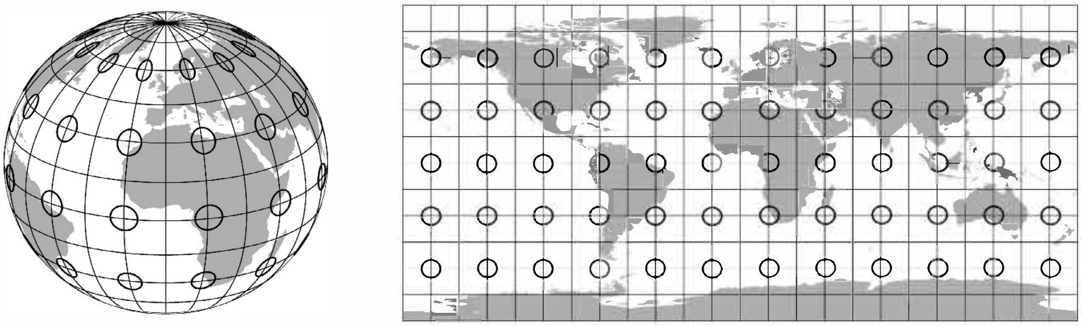


Figure 2-7: Geographic coordinates on a spherical (left) and Cartesian (right) representation. Notice that "circles" defined by a 5 degree radius do not form circles on the Earth's surface near the poles, as shown on the spherical representation (left figure), but appear as circles in the highly distorted Cartesian plot of geographic coordinates (right). This figure illustrates both that a) the surface distance for a unit of longitude changes depending on your location on Earth, and b) a Cartesian plot of geographic coordinates is highly distorted.

torted map. Note the distorted shape of Antarctica in Figure 2-7, right.

Because the spherical system for geographic coordinates is non-Cartesian, formulas for area, distance, angles, and other geometric properties used in a Cartesian coordinate system should not be used with geographic coordinates. Areas are usually calculated after converting to a projected system, described in chapter 3.

There are two primary conventions used for specifying latitude and longitude (Figure 2-8). The first uses a leading letter, N, S, E, or W, to indicate direction, followed by a number to indicate location. Northern latitudes are preceded by an N and southern latitudes by an S, for example, N90°, S10°. Longitude values are preceded by an E or W, for example W110°. Longitudes range from 0 to 180 degrees east or west. Note that the east and west longitudes meet at 180 degrees, so that E180° equals W180°.

Signed coordinates are the second common way to specify latitude and longitude. Northern latitudes are positive and southern latitudes are negative, and eastern longitudes positive and western longitudes negative. Latitudes vary from -90 degrees to 90 degrees, and longitudes vary from -180 degrees to 180 degrees. By this convention, the longitudes "meet" at the maximum and minimum values, so -180° equals 180°.

Coordinates may easily be converted between these two conventions. North latitudes and east longitudes are converted by removing the leading N or E. South latitudes and west longitudes are converted by first removing the leading S or W, and then changing the sign of the remaining number from a positive to a negative value.

Spherical coordinates are most often recorded in a degrees-minutes-seconds (DMS) notation: N43° 35′ 20″ for 43 degrees, 35 minutes, and 20 seconds of lati-

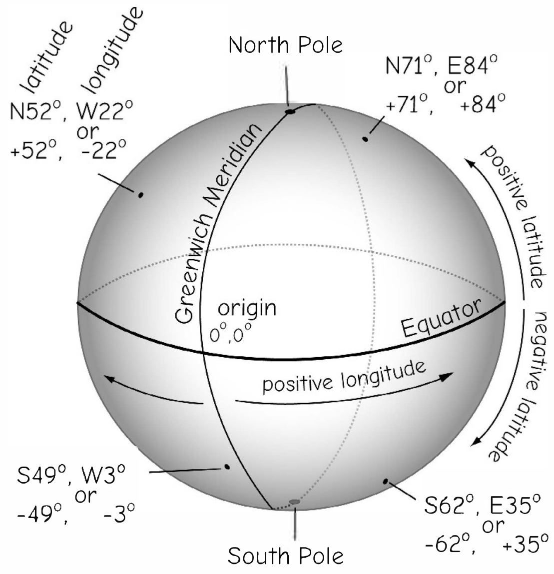


Figure 2-8: Spherical coordinates of latitude and longitude are most often expressed as directional (N/S, E/W), or as signed numbers. Latitudes are positive north, negative south; longitudes are positive east, negative west.

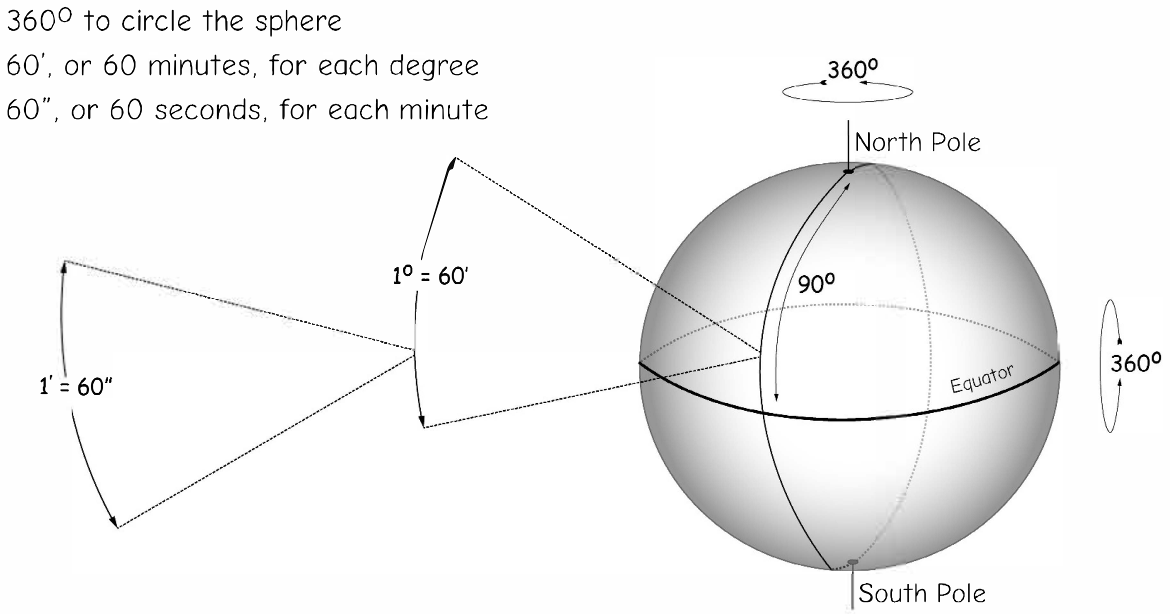


Figure 2-9: There are 360 degrees in a complete circle, with each degree composed of 60 minutes, and each minute composed of 60 seconds.

tude. In DMS, each degree is made up of 60 minutes of arc, and each minute is in turn divided into 60 seconds of arc (Figure 2-9). This yields 60 times 60, or 3600 seconds for each degree of latitude or longitude. Note that the ancient Babylonians established these splits almost 4,000 years ago, defining 360 degrees for a complete circle, and we've carried this convention down to today.

Spherical coordinates may also be expressed as decimal degrees (DD). When using DD, the degrees take the usual -180 to 180 (longitude) and -90 to 90 (latitude) ranges, but minutes and seconds are reported as a decimal portion of a degree (from 0 to 0.99999...).

Conversion between DMS and DD is shown in Figure 2-10.

DD from DMS

DD = D + M/60 + S/3600
e.g.
DMS = 32° 45' 28"

DD = 32 + 45/60 + 28/3600
= 32 + 0.75 + 0.0077778
= 32.7577778

DMS from DD

D = integer part
M = integer of decimal part x 60
S = 2nd decimal x 60
e.g.
DD = 24.93547
D = 24
M = integer of first decimal x 60
= 0.93547 x 60
= integer of 56.1282
= 56
S = 2nd decimal x 60
= 0.1282 * 60 = 7.692
so DMS is
24° 56' 7.692"

Figure 2-10 Examples for converting between DMS and DD expressions of spherical coordinates.

Spherical vs. Ellipsoidal Earth

While we often describe the Earth's shape as a sphere, it is better approximated as an ellipsoid. A sphere is a solid object defined by a center location and an equal radius in all directions. An ellipsoid is an approximately spherical solid, but with unequal radii along the axes. Spheroids and ellipsoids may be viewed in cross-section, revealing their difference in shape (Figure 2-11). The Earth's shape is best viewed as an ellipsoid flattened in the north-south direction. This flattening is quite small, approximately one part in 300. Translated to human scales, this is about an 8 mm (1/30th of an inch) flattening in a basketball. While difficult to observe directly, it is large enough to distort common geodetic measurements and navigation on the surface of the Earth. Many navigation and measurement estimates have two sets of formulas, one an approximation based on a purely spherical globe, and a more complicated and precise set based on an ellipsoidal shape.

Note that the words spheroid and ellipsoid are often used interchangeably. GIS software often prompts the user for a sphere or spheroid when defining a coordinate projection, and then lists a set of ellipsoids, with differing polar and equatorial radii.

The best estimates of Earth's radii, a and b, have evolved as measurement systems have improved. Today, the best estimate for a is 6,378,137.0 meters (m), and for b 6,356,752.3 m. A mean value of 6,367,444.7 m is often used for spheroids, but sometimes the value for a is adopted, 6,378,137 m, or just 6,378 km.

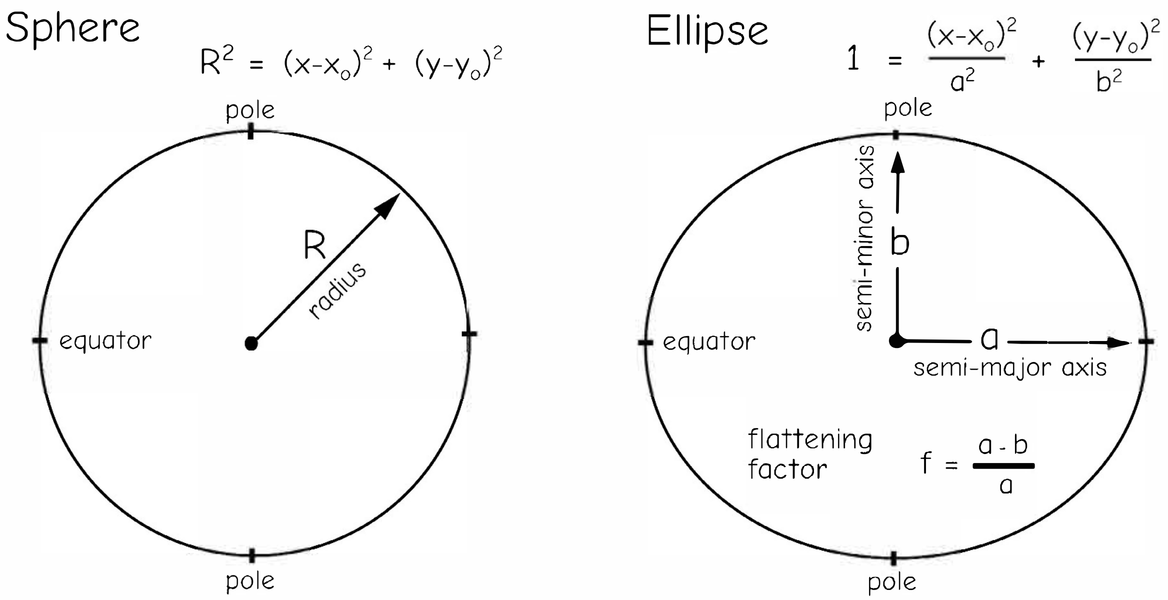


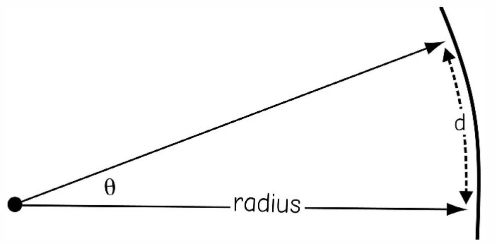
Figure 2-11: Spherical (left) vs. ellipsoidal (right) approximations of the Earth's shape. A sphere has a single radius, while an ellipse has different radii along the semi-major and semi-minor axes. The spheroid and ellipsoid can be thought of as rotating these two basic shapes around the polar axis to create solid figures.

Converting Arc to Surface Distances

At times we need to calculate the distance on the surface of the Earth that is spanned by an arc measure. For example, I might have two locations that differ by 10 seconds of arc, and wish to estimate the distance between them. We can approximate the surface distance on a circle or sphere by the formula:

$$d = r \cdot \theta \tag{2.1}$$

where d is the approximate ground distance, r is the radius of the circle or sphere, and θ is the angle of the arc. There is a more complicated formula for ellipsoidal surfaces, but the above formula is acceptable for most applications.



d = radius • θ

where θ is measured in radians,
with
1 radian = 57.2957°

Given an Earth radius of 6,378,137 m, how much distance is spanned by 10" of arc?

Arc= 10"/3600"/1° = 0.00277778°

=0.00277778°/57.2957 degrees per radian
= 0.000048481435 radians

d = 6378137m • 0.000048481435
= 309.2 meters

Figure 2-12: Example calculation of the approximate surface distance spanned by an arc.

Converting degrees to radians:

30.1487 degrees is

30.1487 / 57.2957795

= 0.52619 radians

Converting radians to degrees:

1.284 radians is

1.284 x 57.2957795

= 73.5678 degrees

Figure 2-13: Conversion between radian and degree angle units.

Figure 2-12 shows an example calculation of arc length, using the average radius for Earth. Note that equation (2.1) applies to a generic arc angle, measured in the direction of the spanned arc, without regard to the latitude/longitude system. Substituting latitude values will result in a reasonably accurate answer, but substituting longitude values anywhere but along the Equator will result in an error, largest near the poles, due to longitudal convergence. The formula is best used as a first approximation of distance spanning generic arcs, and not using longitudinal coordinates.

Note that the angle should be specified in radian measure, defined as 2π radians per the 360 degrees, or approximately 57.2957795 degrees per radian. Radian measures are an alternative to degrees, and scale the rotation by the radius of the circle. You may easily convert between radian and degree units (Figure 2-13). Many spreadsheet, online, and app programs by default use radian measure, and substituting degrees will lead to errors.

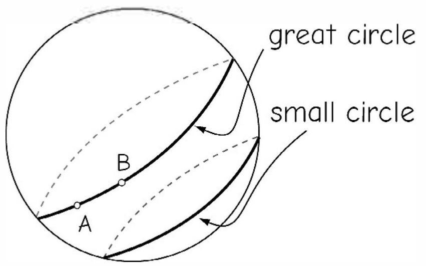
Great Circle Distance

Spherical approximation

Consider two points on the Earth's surface,

A with latitude, longitude of (ϕ_A, λ_A), and

B with latitude, longitude of (ϕ_B, λ_B)



The great circle distance between points on a sphere is given by the formula:

$$d = r \cdot 2\sqrt{\sin^{-1}[(\sin^2(\tfrac{\Delta\phi}{2})) + \cos(\phi_A) \cdot \cos(\phi_B) \cdot \sin^2(\tfrac{\Delta\lambda}{2})]}$$

where d is the shortest distance on the surface of the Earth from A to B, r is the Earth's radius, approximately 6378 km, and $\frac{\Delta\phi}{2}, \frac{\Delta\lambda}{2}$ are the differences between point latitudes and longitudes, divided by two.

As an example, the distance between Paris, France, and Seattle, USA, is:

Latitude, longitude of Paris, France = 48.864716°, 2.349014°

Latitude, longitude of Seattle, USA = 47.655548°, -122.30320°

$$d = 6378 \cdot 2\sqrt{\sin^{-1}[(\sin^2(0.604584)) + \cos(48.864716) \cdot \cos(47.655548) \cdot \sin^2(62.36107)]}$$

$$= 8{,}034.8391\text{km}$$

Figure 2-14: Calculation of the great circle distance between points.

The great circle distance formula should be used to estimate the surface distance between two points when using latitudes/longitudes (Figure 2-14). A *great circle* is defined by any plane that intersects a globe and passes through it's center. The Equator and meridians are great circles, while lines of equal latitude other than the Equator are not great circles. A great circle distance is the shortest path on the Earth's surface between two points, and long-distance airline routes approximate great circles. As with all trigonometric formulas, you should know if your calculations expect degree or radian measures as input, and convert accordingly.

Three-Dimensional, Earth-Centered Coordinates

We noted an alternate, three-dimensional (3-D) Cartesian representation of coordinates for locations, typically in, on, or near the Earth (Figure 2-15). This is commonly used in geodesy, the science of the Earth's shape, size, and physical dynamics, that underpins all coordinate measures. Geodesy is at the heart of map projections (Chapter 3) and satellite positioning (Chapter 5), fundamental building blocks of GIS.

The 3D Cartesian system typically places the origin near or at the mass center of the Earth. This Cartesian system is aligned with the Z axis through the geographic North Pole and the X and Y axes forming a plane

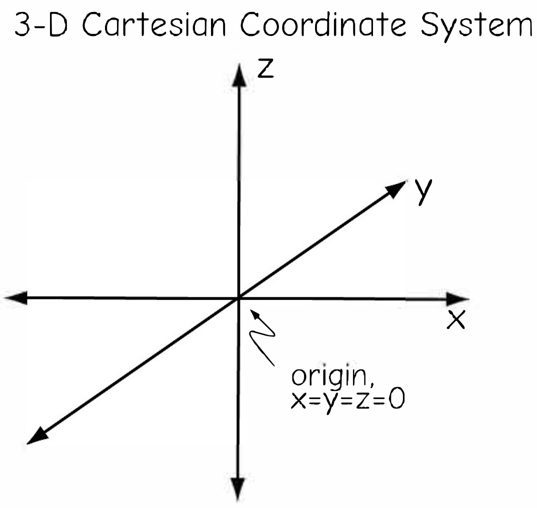


Figure 2-15: A 3-D Cartesian coordinate system.

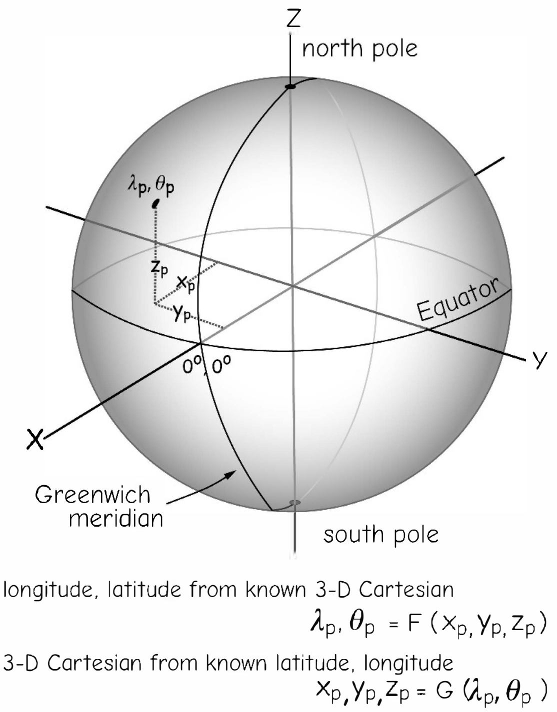


Figure 2-16: Formulas exist to convert between known spherical geographic coordinates (latitude and longitude on a spheroid) and corresponding 3-D Cartesian coordinates (see appendix C).

on the Equator (Figure 2-16). The positive X-axis intersects the ellipsoid where latitude and longitude values are both zero, and the positive Y-axis intersects the ellipsoid at a longitude of 90 and latitude of 0.

Mathematical formulas allow us to calculate any X, Y, and Z given any latitude, longitude, and Earth radii (Figure 2-16). Each latitude/longitude/radius coordinate in the geographic system corresponds to an X-Y-Z triplet in the 3-D Cartesian coordinate system. These formulas are commonly used by geodesists in the most precise surveys, but are also embedded in many softwares that convert between different versions of our coordinate data.

There are two different sets of equations, one assuming a spherical Earth, and a more accurate one assuming an ellipsoidal Earth. A detailed discussion of these is best left for an advanced course, so formulas are included in Appendix C for reference.

Geographic and Magnetic North

There is often confusion between magnetic north and geographic north. Magnetic north and the geographic north do not coincide (Figure 2-17). Magnetic north is the location towards which a compass points. The geographic North Pole is the average northern location of the Earth's axis of rotation. If you were standing on the geographic North Pole with a compass, it would point approximately in the direction of the Bering Straits, and some 200 kilometers away. In addition, Magnetic North "wanders" through time, and has recently increased it's rate of shift (Figure 2-17).

Because magnetic north and the geographic North Pole are not in the same place, a compass does not point towards geographic north when observed from most places on Earth. The compass will usually point east or west of geographic north, defining an angular difference called the magnetic *declination.* Declination varies across the globe, and also has varied through time as magnetic north wanders.

Note that our definition of geographic north is the average northern location of the Earth's axis of rotation. We say average because the Earth wobbles, or nutates, on its axis. This means the axis location varies

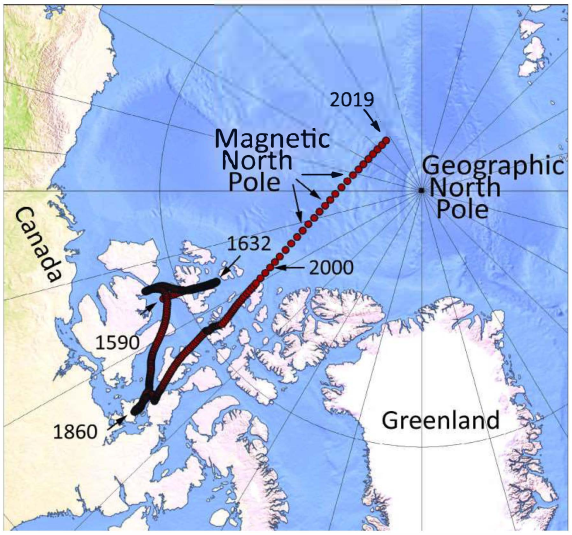


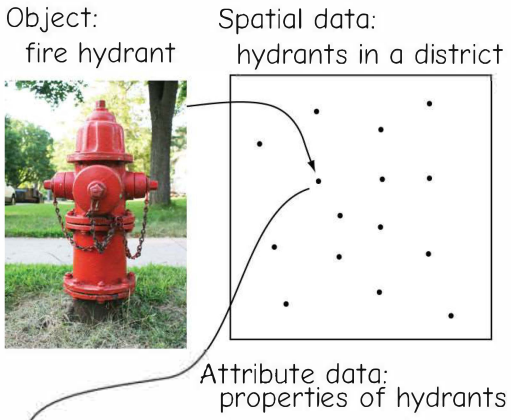
Figure 2-17: Magnetic and geographic North Poles. Year dates show how the Magnetic North has wandered through time, increasing in velocity over the past few decades.

slightly, within a circle about 9 meters (30 feet) across, so the northern pole location is always within this circle. The nutation has a period of 433 days, with the pole returning back to its original location over that time.

Attribute Data and Types

Attribute data are used to record the non-spatial characteristics of an entity. Attributes, also called *items* or *variables*, may be envisioned as a list of characteristics that describe features. Color, depth, weight, owner, vegetation type, or land use are examples of variables that may appear as attributes. Attributes record values; for example, a fire hydrant may be colored red, yellow, or orange, have 1 to 4 flanges, and a pressure rating of any real number from 0 to 12,000.

Attributes are often presented in tables and arranged in rows and columns (Figure 2-18). Each row corresponds to a spatial object, and each column corresponds to an attribute. Tables are often organized and managed using a specialized computer program called a database management system (DBMS, described more fully in Chapter 8).



ID	Flanges	Color	Rating
1	3	red	high
2	3	red	high
3	2	green	low
4	3	red	mid
.	.	.	.

Figure 2-18: Attributes are typically envisioned in a table, with objects arranged in rows and attributes aligned in columns.

All attributes can be categorized as nominal, ordinal, or interval/ratio attributes. *Nominal attributes* are variables that provide descriptive information about an object. The color is recorded for each hydrant in Figure 2-18. Other examples of nominal data are vegetation type, a city name, the owner of a parcel, or soil series. There is no implied order, size, or quantitative information contained in nominal attributes.

Nominal attributes may also be images, film clips, audio recordings, or other descriptive information, for example, GIS for real estate often have images of the buildings as part of the database. Image, video, or sound recordings stored as attributes are sometimes referred to as "BLOBs" for *binary large objects*.

Ordinal attributes imply a ranking by their values. An ordinal attribute may be descriptive, such as high, mid, or low, or it may be numeric; for example, an erosion class with values from 1 to 10. The order reflects only rank, and not scale. An ordinal value of four has a higher rank than two, but we can't infer that the attribute value is twice as large.

Interval/ratio attributes are used for numeric items where both rank order and absolute difference in magnitudes are represented, for example, the number of flanges in the second column of Figure 2-18. These data are often recorded as real numbers on a linear scale. Area, length, weight, height, or depth are a few examples of attributes that are represented by interval/ratio variables.

Items have a *domain*, a range of values they may take. Colors might be restricted to red, yellow, and green; cardinal direction to north, south, east, or west; and size to all positive real numbers.

Common Spatial Data Models

All spatial data models are based on a conceptualization. As an example, consider a regional map that defines roads as lines. We conceive of each road as a linear feature that fits into a small number of categories. These lines connect cities and towns that are shown as discrete points or polygons on the map. Road properties may include only the road type, e.g., highway or local road. The roads have a width represented by a line symbol on the map; however, the scaled road width may not represent the true road width. All state highways are represented equally although they may vary. Some may have wide shoulders, others not, or dividing barriers of concrete, versus a broad vegetated median, but we may choose to omit this variation, fitting all highways into one class.

There are two main conceptualizations used for digital spatial data. The first defines discrete objects using a *vector data model.* This model uses discrete elements such as points, lines, and polygons to represent the geometry of real-world entities (Figure 2-19, left).

Farm fields, roads, wetlands, cities, and census tracts are examples of entities that are often represented by discrete vector objects. Points are often used to define the locations of "small" objects such as wells, buildings, or ponds. Lines may be used to represent linear objects, for example, rivers or roads, or to enclose polygons, which identify area objects. Starting points and ending points for a line are sometimes referred to as *nodes*, while intermediate points in a line are referred to as *vertices*.

Vector objects are discrete. A forest may share an edge with a pasture, and this boundary is represented by lines. In truth, a forest edge may grade into a mix of trees and shrubs, then shrubs and grass, then pure grass; however, in the vector conceptualization, a line between two land cover types will be drawn to indicate a discrete, abrupt transition. Lines and points have coordinate locations, but points have no dimension, and lines have no dimension perpendicular to their direction. Area features are defined by a closed, connected set of lines.

The second common conceptualization identifies and represents grid cells for a given region of interest. This conceptualization employs a *raster* data model (Figure 2-19, right). Raster cells are arrayed in a row and column pattern to provide "wall-to-wall" coverage of a study region. Cell values are used to represent the type or quality of

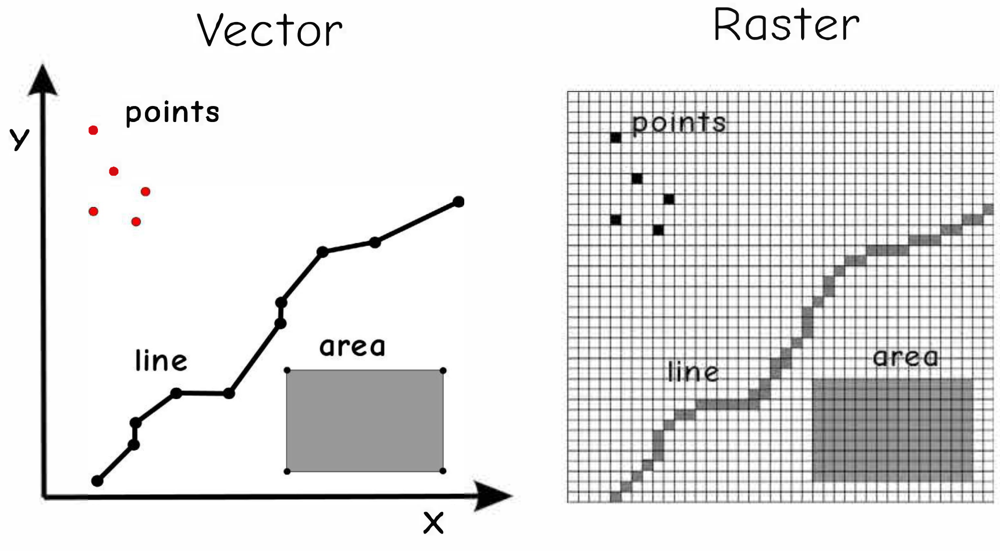


Figure 2-19: Vector and raster data models.

mapped variables. Raster models are often used with variables that may change continuously across a region. Elevation, mean temperature, slope, average rainfall, cumulative ozone exposure, or soil moisture are examples of phenomena that are often represented as continuous fields. Raster representations are also sometimes used to represent discrete features, for example, class maps of vegetation or political units.

Data models are often interchangeable in that many phenomena may be represented by many data models. For example, elevation may be represented as a raster surface (continuous field) or as a series of lines representing contours of equal elevation (discrete objects). Data may be converted from one model to another; for example, the location of contour lines may be determined by evaluating the raster surface, or a raster data layer may be derived from a set of contour lines. These conversions entail some costs both computationally and perhaps in data accuracy.

The decision to use either a raster or vector model often depends on our conceptualization of the objects and the most frequent operations performed. We think of elevation as a continuous variable and slope is more easily determined when elevation is represented in a raster data set. However, discrete contours are often the preferred format for printed maps, so the discrete conceptualization of a vector data model may be preferred in some cases. The best data model for a given application depends on the most common operations, the experiences and views of the GIS users, the form of available data, and the influence of the data model on data quality.

Other, less common data models are sometimes used. A triangulated irregular network (TIN) is one such model, employed to represent surfaces such as elevations, through a combination of point, line, and area features. We will introduce and discuss less common data models later in this chapter.

Vector Data Models

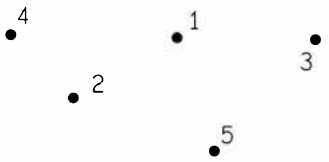
A vector data model uses sets of coordinates and associated attribute data to define discrete objects. Groups of coordinates define the location and boundaries of discrete objects, and these coordinate data plus their associated attributes are used to create vector objects representing the real-world entities (Figure 2-20). In the most common vector models, there is an attribute table associated with each vector layer, and a single row in the table corresponding to each feature in the data layer. These vector layers are said to contain *single-part features*, because there is a single geographic object for each row in the table, with one to several columns in each row. All values in a column have the same type, so for any given column, all entries might be ordinal, or interval ratio, or a BLOB, or some other defined type. An identifier value, or ID, is typically included, and this value is often unique within the table, with an unrepeated value assigned for each row and corresponding feature.

There are three basic types of vector objects: points, lines, and polygons (Figure 2-20, top). A point uses a single coordinate pair to represent the location of an entity that is considered to have no dimension. Gas wells, light poles, accident location, and survey points are examples of entities often represented as point objects. Some of these have real physical dimension, but for the purposes of the GIS users they may be represented as points. In effect, this means the size or dimension of the entity is not important, only its location.

Attribute data are attached to each point, and these attribute data record the important non-spatial characteristics of the point entities (Figure 2-20). When using a point to represent a light pole, important attribute information might be the height of the pole, the type of light and power source, and the last date the pole was serviced.

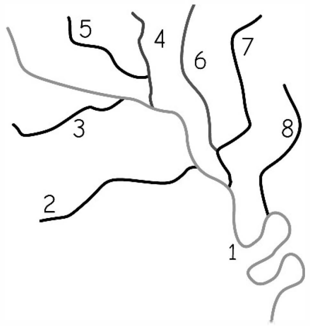
Linear features are represented as lines in vector data models (Figure 2-20, mid). Lines are most often represented as an ordered set of coordinate pairs. Each line is

Points



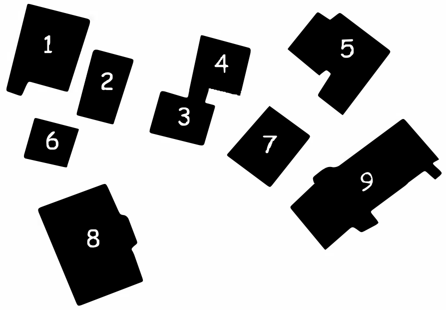
ID	Tower Name	Height	Format
1	WKRP	101.0	Pop
2	WYOU	55.5	Oldies
3	TPT	486.0	Public TV
4	WQXR	99.5	Classical
5	BBC	212.1	News

Lines



ID	Name
1	Tuckaseegee River
2	Pigeon Branch
3	Poplar Run
4	Shope Fork
5	Mel's Brook
6	Merdesansrame Creek
7	Longue Arm
8	Arroyo Grande

Polygons



ID	Building Name	Floors	Roof Type
1	Hodson Hall	6.0	flat, sealed tar
2	Borlaug Hall	5.5	pitched 9/12, tile
3	Guilford Technology Bldg.	4.0	flat, gasket
4	Shop Annex	2.5	flat, sealed tar
5	Animal Sciences Bldg.	1.0	pitched 12/12, tile
6	Administration Bldg.	14.0	pitched 6/12, metal
7	Climate Sciences Center	6.0	flat, sealed tar
8	Grantham Tower	1.0	pitched 9/12, tile
9	Biological Sciences Bldg.	9.0	pitched 12/12, tile

Figure 2-20: An example of the most common vector data model structures. Geographic features consist of points, lines, or polygons, with each feature corresponding to a row in a table with an identifier (ID) and a set of attributes arrayed in columns.

made up of line segments that run between adjacent coordinates in the ordered set. Attributes in a table correspond to line segments (Figure 2-20, mid). Curved linear entities are most often represented as a collection of short, straight, line segments, although curved lines are at times represented by a mathematical equation describing a geometric shape. The line starting and ending points are often called nodes, and intermediate points used to represent the line shape are called vertices.

Area entities are most often represented by closed polygons (Figure 2-20, bottom). These polygons are formed by a set of connected lines, either one line with an ending point that connects back to the starting point, or as a set of lines connected start-to-end. Polygons have an interior region and may entirely enclose other polygons in this region. Polygons may be adjacent to other polygons and thus share “bordering” or “edge” lines with other polygons. Attribute data such as area, perimeter, land cover type, or county name may be linked to each polygon (Figure 2-20, bottom).

Note that there is no uniformly superior way to represent features, and we may represent the same features as points, lines, or polygons (Figure 2-21). Some feature types may appear to be more "naturally" represented one way: manhole covers as points, roads as lines, and parks as polygons. However, in a very detailed data set, the manhole covers may be represented as circles, and both edges of the roads may be drawn and the roads represented as polygons. The best representation depends on the detail, accuracy, and intended use of the data set.

Vector layers sometimes have a many-to-one relationship between geographic features and table rows (Figure 2-22), defining *multi-part features*. In these instances, many spatially distinct features are matched with a row, and the row attributes apply to all the distinct features. This is common when representing islands, groups of buildings, or other clusters of features that make up a perceived whole thing. These multi-part features may have multiple geographic objects that correspond to one row.

Multi-part features may also be used for large data sets, for example, when millions of point observations are collected automatically with laser scanners. Tables are often slower to process than point geographies, and so reducing the table size by grouping points into multi-part features may shorten many operations.

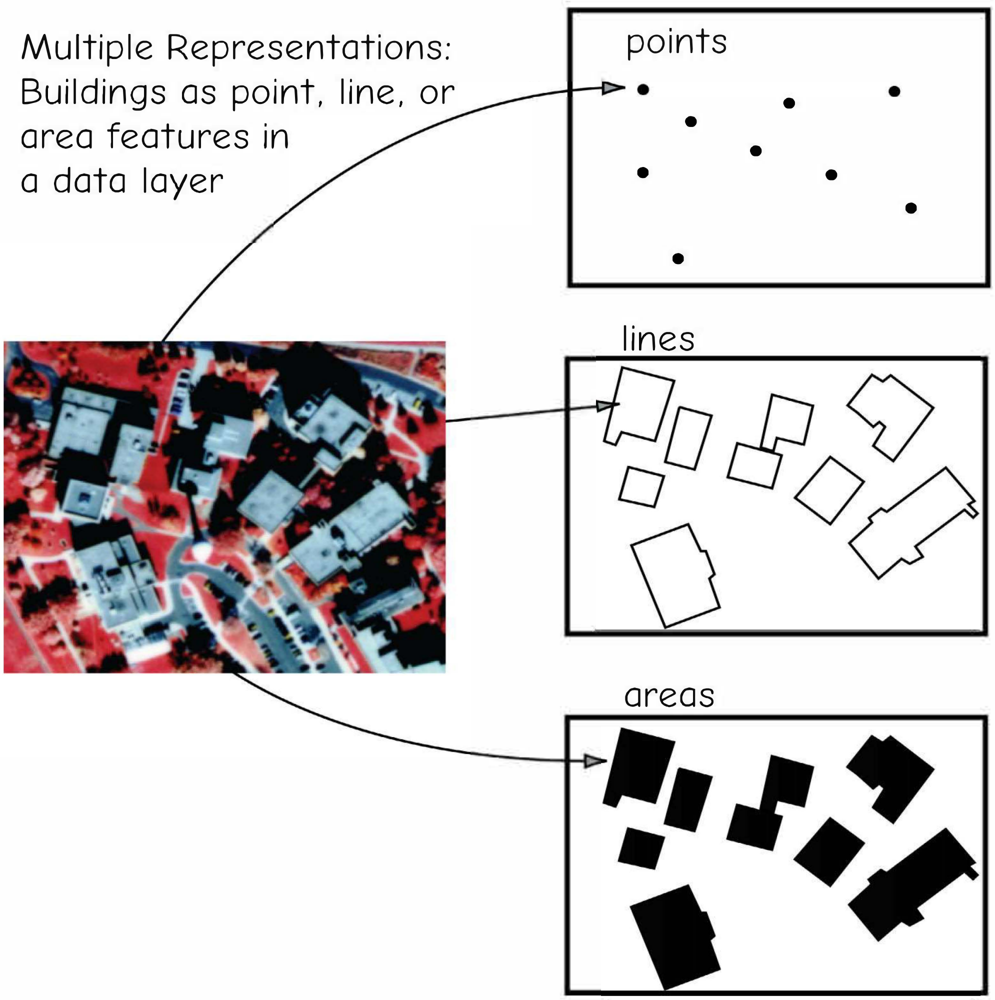
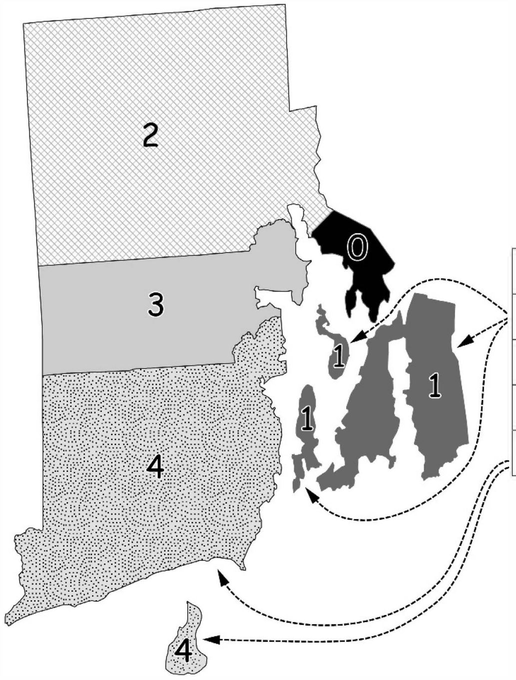
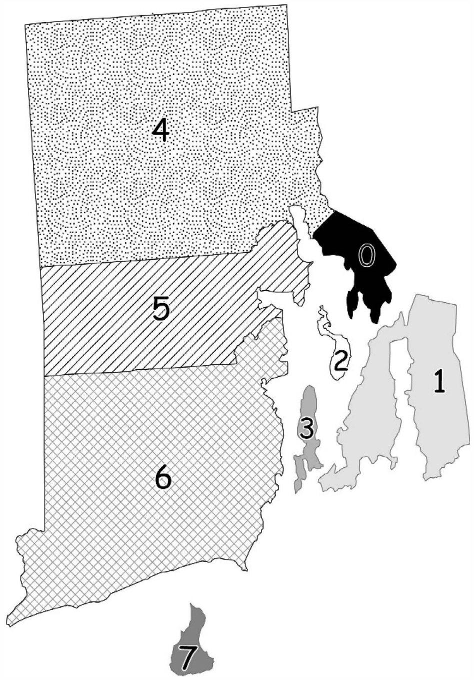


Figure 2-21: The same objects may be represented by points, lines, or polygons, depending on our view of and intended use for the data.



Multi-part Features

ID	County	Area
0	Bristol	73.5
1	Newport	286.5
2	Providence	1,121.9
3	Kent	455.6
4	Washington	912.3



Single-part Features

ID	County	Area
0	Bristol	73.5
1	Newport	243.5
2	Newport	14.8
3	Newport	24.8
4	Providence	1,121.9
5	Kent	912.3
6	Washington	884.7
7	Washington	27.6

Figure 2-22: Example of multi-part and single-part features. Here, counties for Rhode Island, a state in the Eastern U.S.A., are shown with one table entry for each county (top), with multi-part features in a layer, and with one table entry for each distinct polygon (bottom), with only single-part features in the layer. Note that calculations, analysis, and interpretation may differ for multi-part v.s. single-part features.

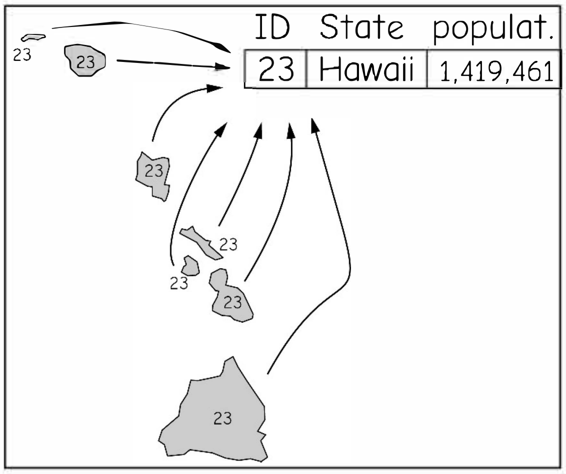
Care is warranted when converting multi-part features to single-part features. The most common problems arise for aggregate variables in polygon layers, such as total counts. For example, population data are often delivered by census areas such as states. Many states, e.g., Hawaii, have several parts and are represented by a multi-part shape. The population is associated with the aggregated set of polygons comprising the state (Figure 2-23). When converted to single-part shapes, the attributes are often copied for each component polygon. In our example, all single-part polygons will be assigned the attribute values for the multi-part feature, in effect repeating counts for each part. Subsequent aggregation or calculation across the population column may result in error.

Attributes for converted shapes may be corrected. If component data are available, they can be assigned to each of the single-part features. If not, then some weighting scheme may be available, for example, if there is a correlation between area and count. Until they are reviewed and appropriately adjusted, single-part attributes derived from multi-part features should be used with caution.

Polygon Inclusions and Boundary Generalization

Vector data frequently exhibit two characteristics: polygon inclusions and boundary generalization. These characteristics are oft-ignored, but may affect the use of vector data. These concepts must be understood, their presence evaluated, and effects weighed in the use of vector data sets.

Polygon inclusions are areas in a polygon that are different from the rest of the polygon, but still part of it. Inclusions occur because we typically assume an area represented by a polygon is homogeneous, but this is often untrue, as illustrated in Figure 2-24. The figure shows a vector polygon layer representing raised landscaping beds (a). The general attributes for the polygon may be coded; for example, the surface type may be recorded as cedar mulch. The area noted in Figure 2-24b shows a walkway that is an inclusion in a raised bed. This walkway has a concrete surface. Hence, this walkway is an unresolved inclusion within the polygon.



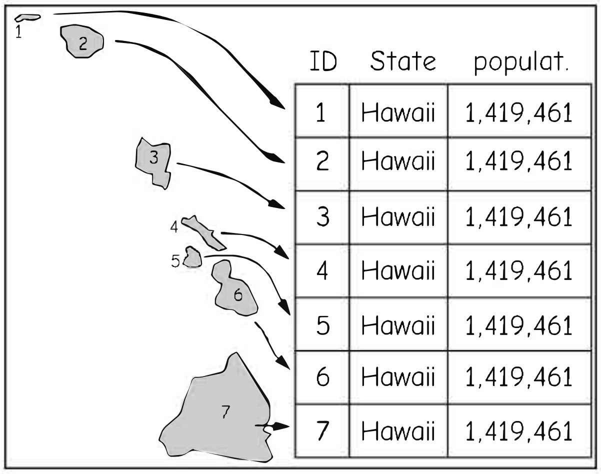


Figure 2-23: multi-part to single-part conversion may lead to errors in subsequent analysis because attributes may be copied from the original, multi-part cluster (left, above), to each single-part component (right, above). Density, sums, or other derived variables often should be re-calculated for single-part features, but often are not, resulting in errors.

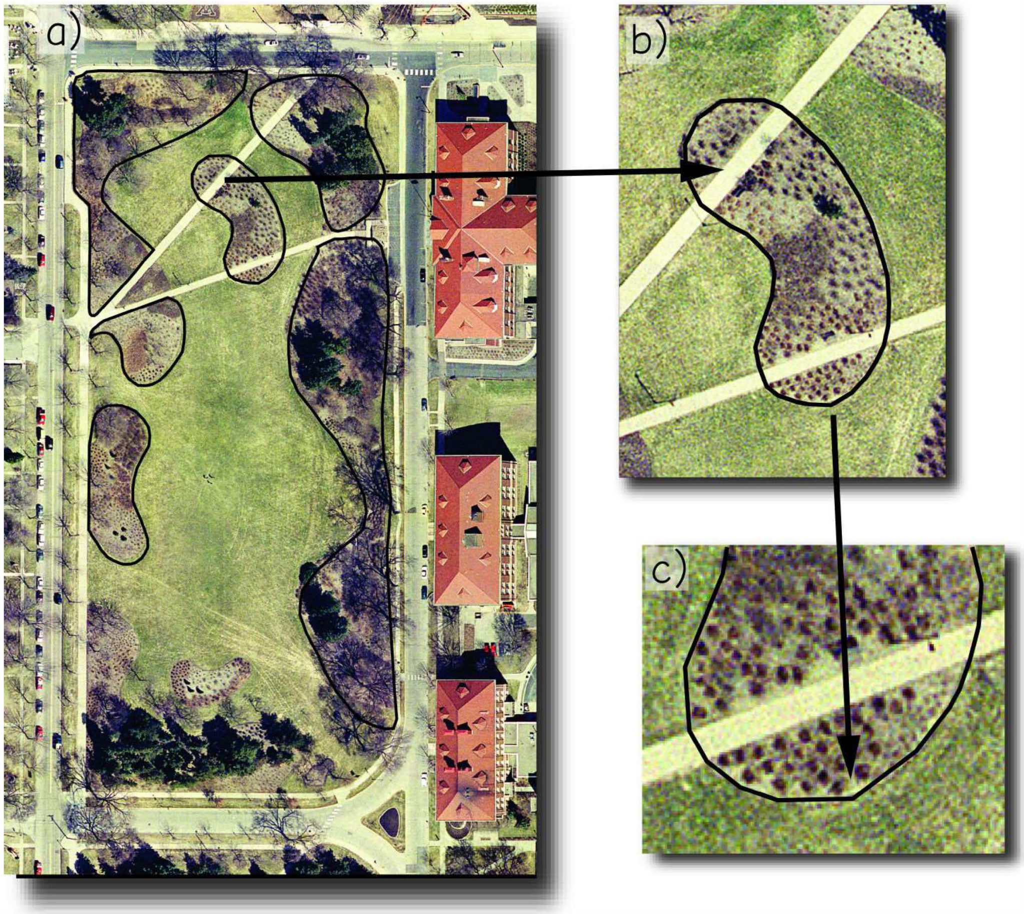


Figure 2-24: Examples of polygon inclusions (sidewalk inclusion in flower bed shown in a and b), and boundary generalization (c) in a vector data model. These approximations typically occur as a consequence of adopting a vector representation, and their impacts must be considered when using vector data.

One solution creates a polygon for each inclusion. This often is not done because it may take too much effort to identify and collect the boundary location of each inclusion, and there typically is some lower limit, or *minimum mapping unit*, on the size of objects we care to record in our data. Inclusions are present in some form in many polygon data layers.

Boundary generalization is the incomplete representation of boundary locations. This problem stems from the typical way we represent linear and area features in vector data sets. As shown in Figure 2-24c, polygon boundaries are represented as a set of connected straight-line segments. The segments are a means to trace the boundaries separating different area features. For curved lines, these straight line segments may be viewed as a sampling of the true curve, and there is typically some deviation of the line segment from the "true" curved boundary. The amount of generalization depends on many factors, and should be so small as to be unimportant for any intended use of the spatial data. However, since many data sets may have unforeseen uses or may be obtained from a third party, the boundary generalization should be recognized and evaluated relative to the specific requirements of any given spatial analysis. There are additional forms of generalization in spatial data, and these are described more thoroughly in Chapter 4.

Vector Topology

Vector data often contain *vector topology,* enforcing strict connectivity and recording adjacency, and planarity. Early systems employed a spaghetti data model (Figure 2-25a), in which lines may not intersect when they should, and may overlap without connecting. The spaghetti model severely limits spatial data analysis and is little used except for very basic data entry or translation. Topological models (Figure 2-25b) create an intersection and place a node at each line crossing, record connectivity and adjacency, and maintain information on the relationships between and among points, lines, and polygons in spatial data. This greatly improves the speed, accuracy, and utility of many spatial data operations.

Topological properties are conserved when converting vector data among common coordinate systems, a common practice in GIS analysis (described in Chapter 3). Polygon adjacency is an example of a topologically invariant property, because the list of neighbors for any given polygon does not change during geometric stretching or bending (Figure 2-25, b and c). These relationships may be recorded separately from the coordinate data.

Topological vector models may vary, and enforce particular types of topological relationships. *Planar topology* requires that all features occur on a two-dimensional surface. There can be no overlaps among lines or polygons in the same layer (Figure 2-26). When planar topology is enforced, lines may not cross over or under other lines. At each line crossing there must be an intersection.

The left side of Figure 2-26 shows nonplanar graphs. In the top left figure, four line segments coincide. At some locations the lines intersect at a node, shown as white-filled circles, but at some locations a line passes over or under another line segment. These lines are nonplanar. The top right of Figure 2-26 shows planar topology enforced for these same four line segments. Nodes are found at each line crossing.

Polygons can also be nonplanar, as shown at the bottom left of Figure 2-26. Two polygons overlap slightly at an edge. This may be due to an error; for example, the two polygons share a boundary but have been recorded with an overlap, or there may be two areas that overlap in some way. If topological planarity is enforced, these two polygons must be resolved into three separate, nonoverlapping polygons. Nodes are placed at the intersections of the polygon boundaries (lower right, Figure 2-26).

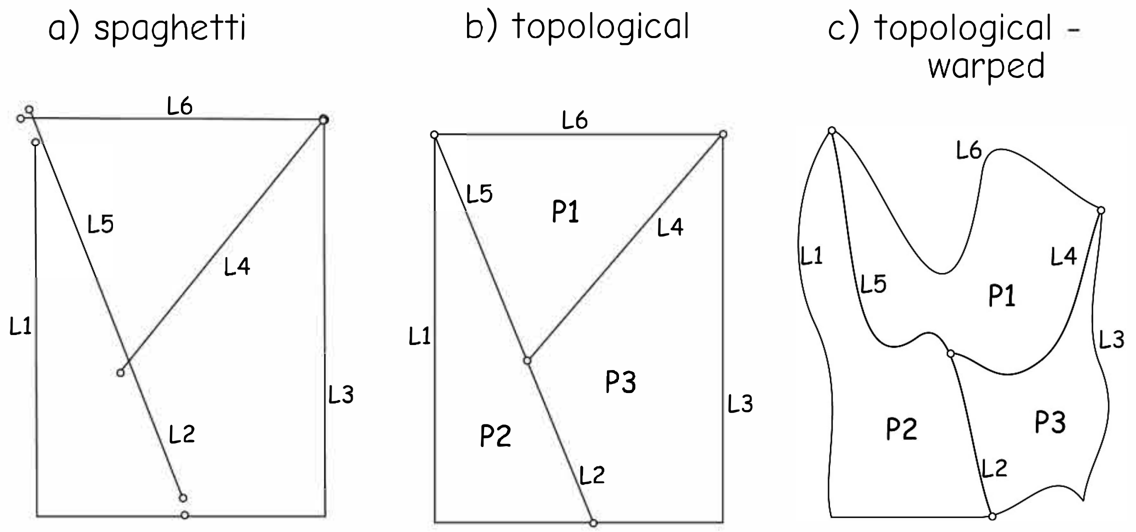


Figure 2-25: Spaghetti (a), topological (b), and topological warped (c) vector data. Figures b and c are topologically identical because they have the same connectivity and adjacency.

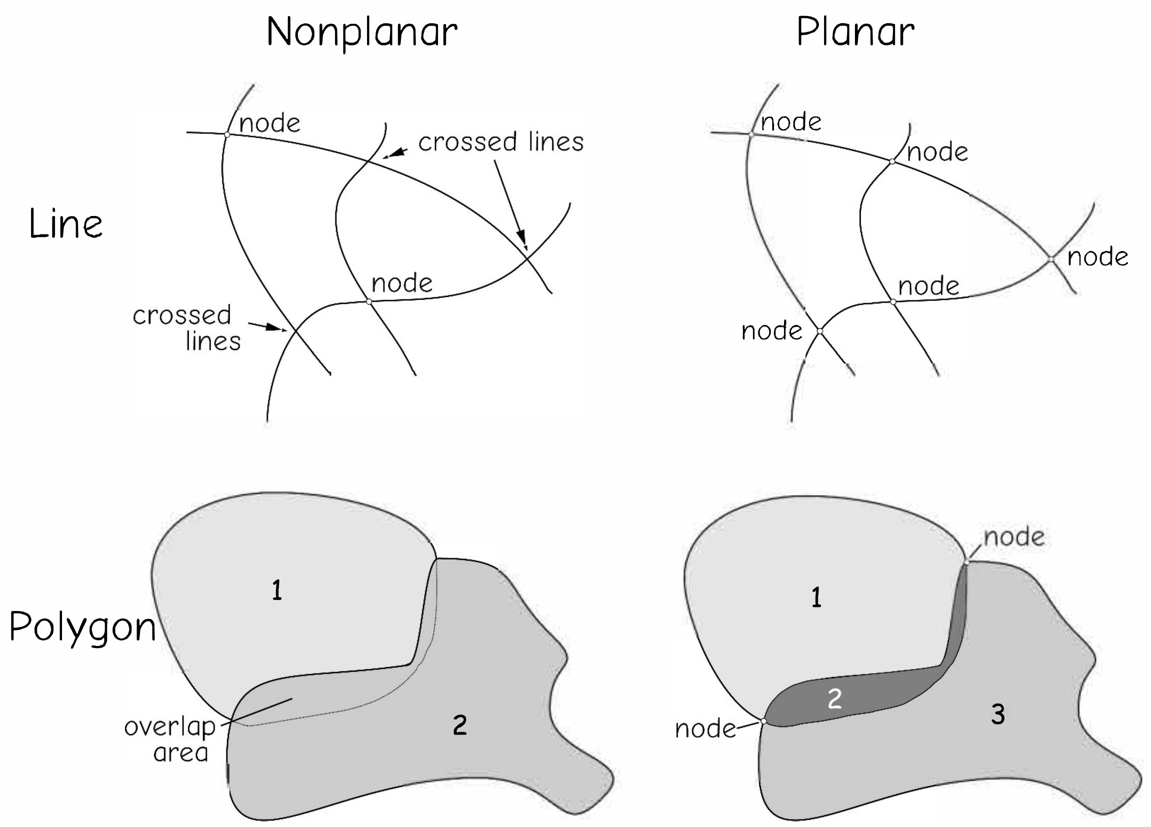


Figure 2-26: Nonplanar and planar topology in lines and polygons.

There are additional topological constructs besides planarity that may be specified. For example, polygons may be exhaustive, in that there are no gaps, holes, or "islands" allowed. Line direction may be recorded, so that a "from" and "to" node are identified in each line. Directionality aids the representation of river or street networks, where there may be a natural flow direction.

There is no uniform set of topological relationships that are included in all topological data models. Different vendors have incorporated different topological information in their data structures. Planar topology is often included, as are representations of *adjacency* (which polygons are next to which) and *connectivity* (which lines connect to which).

Some GIS software create and maintain detailed topological relationships in their data. This results in more complex and perhaps larger data structures, but access is often faster, and topology provides more consistent, "cleaner" data. Other systems maintain little topological information in the data structures, but compute and act upon topology as needed during specific processing.

Topology may also be specified between layers, because we may wish to enforce spatial relationships between entities that are stored separately. As an example, consider a data layer that stores property lines (cadastral data), and a housing data layer that stores building footprints (Figure 2-27). Rules may be specified that prevent polygons in the housing data layer from crossing property lines in the cadastral data layer. This would indicate a building that crosses a property line. Most such instances occur as a result of small errors in data entry or misalignment among data layers. Topological restrictions between two data layers avoid these inconsistencies. Exceptions may be granted in those few cases when a building truly does cross property lines.

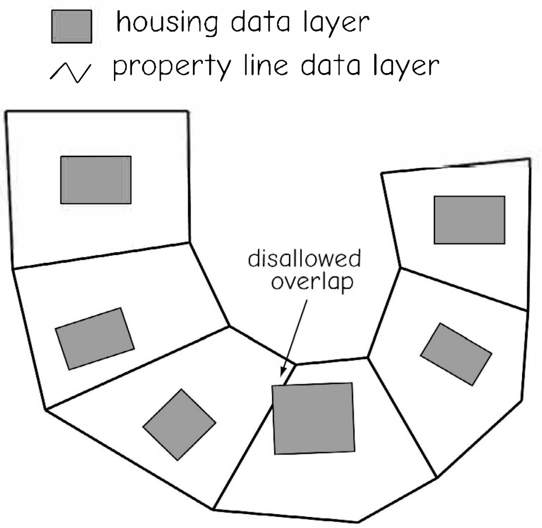


Figure 2-27: Topological rules may be enforced across data layers. Here, rules may be specified to avoid overlap between objects in different layers.

There are many other types of topological constraints that may be enforced, both within and between layers. *Dangles*, lines that do not connect to other lines, may be proscribed, or limited to be greater or less than some threshold length. Lines and points may be required to coincide, for example, water pumps as points in one data layer and water pipes as lines in another, or lines in separate layers may be required to intersect or be coincident. While these topological rules add complexity to vector data sets, they may also improve the logical consistency and value of these data.

Topological vector models often use codes and tables to record topology. As described above, nodes are the starting and ending points of lines. Each node and line is given a unique identifier. Sequences of nodes and lines are recorded as a list of identifiers, and point, line, and polygon topology recorded in a set of tables. The vector features and tables in Figure 2-28 illustrate one form of this topological coding.

Many GIS software systems are written such that the topological coding is not visible to users, nor directly accessible by them. Tools are provided to ensure the topology is created and maintained, that is, there may be directives that require that polygons in two layers do not overlap, or to ensure planarity for all line crossings. However, the topological tables these commands build are often quite large, complex, and linked in an obscure way, and therefore hidden from users.

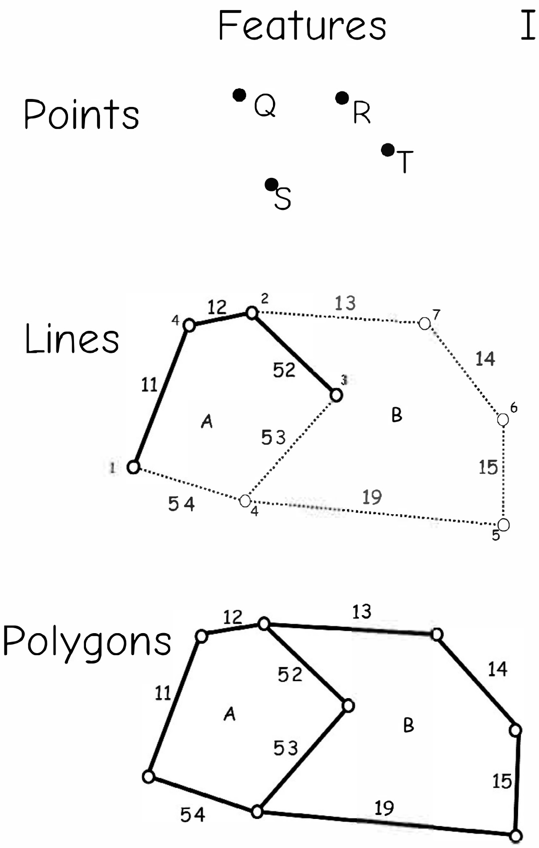
Point topology is often quite simple. Points are typically independent of each other, so they may be recorded as individual identifiers, perhaps with coordinates included, and in no particular order (Figure 2-28, top).

Line topology typically includes substantial structure and identifies at a minimum the beginning and ending points of each line (Figure 2-28, middle). Topology may be organized in tables, including line identifiers, starting nodes, and ending nodes for lines. Lines may be assigned a direction, and the polygons to the left and right of the lines recorded.

Polygon topology may also be defined by tables (Figure 2-28, bottom). The tables may record the polygon identifiers and the ordered list of connected lines that define the polygon.The lines for a polygon form a closed loop, so the starting node of the first line in the list also serves as the ending node for the last line in the list.

Topological models greatly enhance many vector operations. Adjacency analyses are reduced to a "table look-up", a quick and easy operation in most software systems. Assume the city is represented as a single polygon, and we seek all neighboring polygons. Adjacency analysis reduces to 1) scanning the polygon topology table to find the city polygon and reading the list of lines that bound the polygon, and 2) scanning this list of lines, accumulating a list of all left and right polygons. Polygons adjacent to the city may be identified from this list. List searches on topological tables are typically much faster than searches involving coordinate data.

Topological data models often have an advantage of smaller file sizes, largely because coordinate data are recorded once. For example, a nontopological approach



Index and Topology Tables

Point ID	*X*	*Y*
Q	*32.7*	*45.6*
R	*76.3*	*19.5*
S	*22.7*	*15.8*
...		

Line ID	*Begin node*	*End node*	*Left poly*	*Right poly*
11	*1*	*4*	*...*	*A*
12	*4*	*2*	*...*	*A*
52	*2*	*3*	*B*	*A*
...				

Polygon ID	*Lines*
A	*11,12,52,53,54*
B	*52,53,19,15,14,13*
...	

Figure 2-28: An example of vector features and corresponding topology tables. Information on the adjacency, connectivity, and other spatial relationships may be stored in topology tables, and joined to features by indices, here represented by values in the ID columns.

often stores polygon boundaries twice. Lines 52 and 53 at the bottom of Figure 2-28 will be recorded for both polygon A and polygon B. Long, complex boundaries in polygon data sets may double their size. This increases both storage requirements and processing.

There are limitations and disadvantages to topological vector models. First, there are computational costs in defining the topological structure of a vector data layer. Software must determine the connectivity and adjacency information, assign codes, and build the topological tables. Computational costs are typically quite modest with current computer technologies.

Second, the data must be very "clean", in that all lines must begin and end with a node, all lines must connect correctly, and all polygons must be closed. Unconnected lines or unclosed polygons will cause errors during analyses. Significant human effort may be required to ensure clean vector data because each line and polygon must be checked. Software may help by flagging or fixing "dangling" nodes that do not connect to other nodes, and by automatically identifying all polygons. Each dangling node and polygon may then be checked, and edited as needed to correct errors.

Limitations and the extra editing are far outweighed by the gains in efficiency and analytical capabilities provided by topological vector models. Many current vector GIS packages use topological vector models in some form.

Vector Features, Tables, and Structures

As described earlier, geographic features are associated with nonspatial attributes in vector models; tables are used to organize the attributes. In most GIS software, we can most easily view the tables and a graphic representation of the spatial data as a linked table and digital map (Figure 2-29, top).

Most GIS employ underlying file structures to organize components of the spatial data. An example organization is shown in the bottom half of Figure 2-29, where the topological elements are recorded in a linked set of tables, in this example one for each of the polygons, lines, and nodes and vertices. Most GIS maintain the spatial and topological data as a single or cluster of linked files. This internal file structure is often insulated from direct manipulation by the GIS user, but underlies nearly all spatial data manipulations. A user may directly edit or otherwise manipulate table values, usually with the exception of the ID, and the underlying topology and coordinate data are accessed via requests to display, change, or analyze the spatial data components. Data layers may also include additional information (not shown) on the origin, region covered, date of creation, edit history, coordinate system, or other characteristics of a data set.

Note that not all GIS store coordinate and topological data in non-tabular file structures. Coordinates, points, lines, polygons, and other composite features may be stored in tables similar to attribute tables. It is premature to discuss the details of these *spatially enabled* databases, because they are based on something called a *relational data model*, described in detail in Chapter 8. Faster computers support this generally more flexible approach, allowing simpler and more transparent access across different types of GIS software.

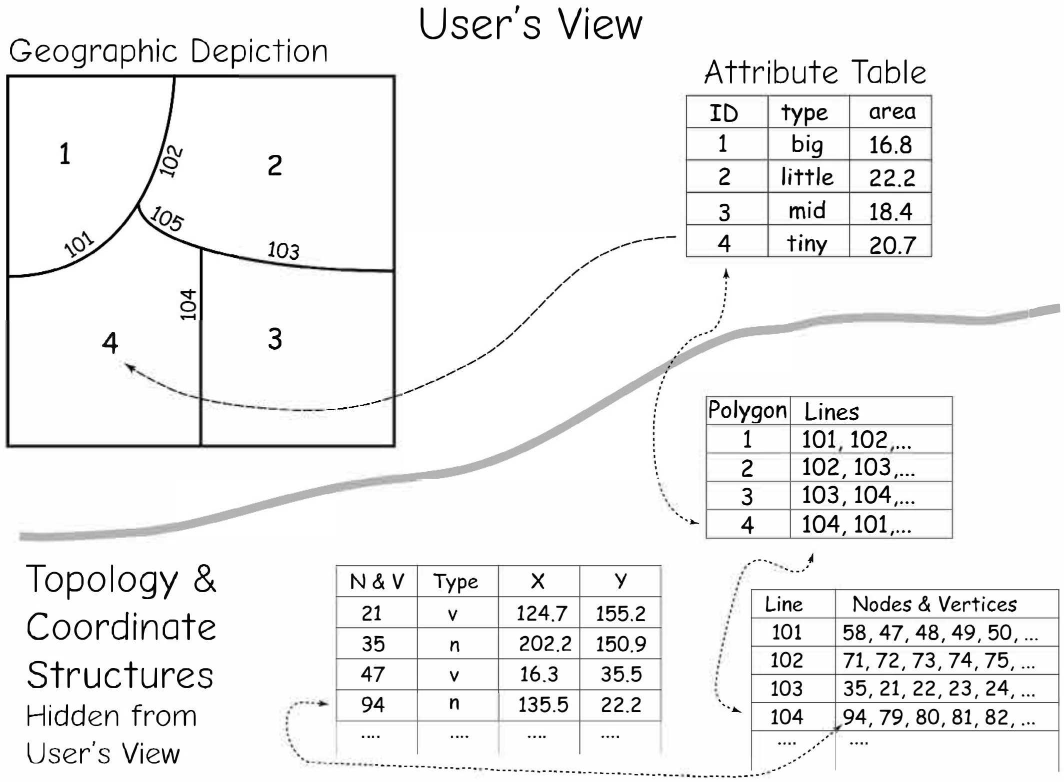


Figure 2-29: Features in a topological data layer typically have a one-to-one relationship with entries in an associated attribute table. The attribute table typically contains a column with a unique identifier, or ID, for each feature. Topology and coordinate data are often hidden from the user, but linked to the attribute and geographic features through pointers and index variables, described in the Data and File Structures section, later in this chapter.

Raster Data Models

Models and Cells

Raster data models define the world as a regular set of cells in a grid pattern (Figure 2-30). Typically, these cells are square and evenly spaced in the x and y directions. The phenomena or entities of interest are represented by attribute values associated with each cell location.

Raster data models are the natural means to represent "continuous" spatial features or phenomena. Elevation, precipitation, slope, and pollutant concentration are examples of continuous spatial variables. These variables characteristically show significant changes in value over broad areas. The gradients can be quite steep (e.g., at cliffs), gentle (long, sloping ridges), or quite variable (rolling hills). Raster data models depict these gradients by changes in the values associated with each cell.

Raster data sets have a *cell dimension*, defining the edge length for each square cell (Figure 2-30). For example, the cell dimension may be specified as a square 30 meters on each side. The cells are usually oriented parallel to the x and y directions, and the coordinates of a corner location are specified.

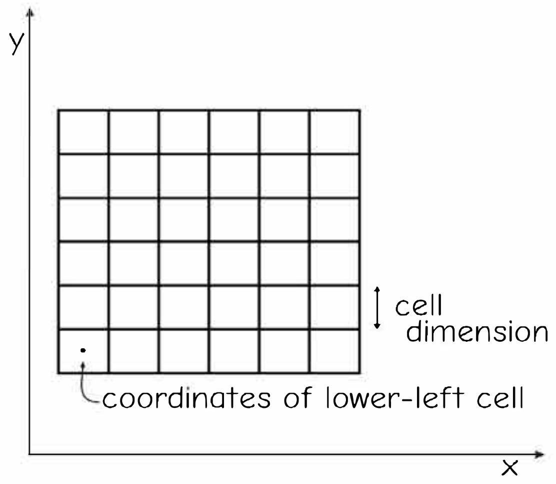


Figure 2-30: Important defining characteristics of a raster data model.

When the cells are square and aligned with the coordinate axes, the calculation of a cell location is a simple process of counting and multiplication. A cell location may be calculated from the cell size, known corner coordinates, and cell row and column number. For example, if we know the lower-left cell coordinate, all other cell coordinates may be determined by the formulas:

$$N_{cell} = N_{low\text{-}left} + row * cell\ size \quad (2.2)$$

$$E_{cell} = E_{low\text{-}left} + column * cell\ size \quad (2.3)$$

where N is the coordinate in the north direction (y), E is the coordinate in the east direction (x), and the row and column are counted starting with zero from the lower left cell.

There is often a trade-off between spatial detail and data volume in raster data sets. The number of cells needed to cover a given area increases four times when the cell size is cut in half (Figure 2-31). Smaller cells provide greater spatial detail, but at the cost of larger data sets.

The cell dimension also affects the spatial precision of the data set, and hence positional accuracy. The cell coordinate is usually defined at a point in the center of the cell. The coordinate applies to the entire area covered by the cell. Positional accuracy is typically expected to be no better than approximately one-half the cell size. No matter the true location of a feature, coordinates are truncated or rounded up to the nearest cell center coordinate. Thus, the cell size should be no more than twice the desired accuracy and precision for the data layer represented in the raster, and often it is specified to be smaller.

Each raster cell represents a given area on the ground and is assigned a value that

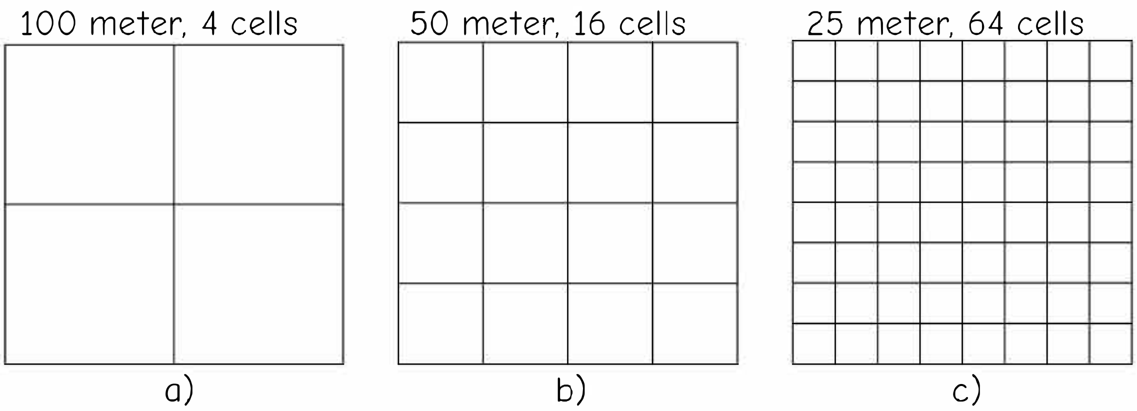


Figure 2-31: The number of cells in a raster data set depends on the cell size. For a given area, a linear decrease in cell size causes an exponential increase in cell number, e.g., halving the cell size causes a four fold increase in cell number.

may be considered to apply to the entire cell. If the variable is uniform across the raster cell, the value will be correct over the cell. However, under most conditions there is within-cell variation, and the raster cell value represents the average, central, or most common value found in the cell. Consider a raster data set representing average weekly income with a cell dimension that is 300 meters (980 feet) on a side. Further suppose that there is a raster cell with a value of 710. The entire 300 by 300 meters area is considered to have this value of 710 pesos per week. There may be many households within the raster cell that do not earn exactly 710 pesos per week. However, the 710 pesos may be the average, the highest point, or some other representative value for the area covered by the cell. While raster cells often represent the average or the value measured at the center of the cell, they may also represent the median, maximum, or another statistic for the cell area.

An alternative interpretation of the raster cell applies the value to the central point of the cell. Consider a raster grid containing elevation values. Cells may be specified as 200 meters square, and an elevation value assigned to each square. A cell with a value of 8,000 meters (26,200 feet) may be assumed to have that value at the center of the cell, but this value will not be assumed to apply to the entire cell.

A raster data model may also be used to represent discrete data (Figure 2-32), for example, to represent land cover in an area. Raster cells typically hold numeric or single-letter alphabetic characters. A coding scheme defines what land cover type the discrete values signify. Each code may be found at many raster cells.

Raster cell values may be assigned and interpreted in at least seven different ways (Table 2-1). We have described three: a raster cell as a point physical value (elevation), as a statistical value (average income), and as discrete data (land cover). Raster values may also be used to represent points and

a	a	a	a	r	f	f	a	a	a	a	a
a	a	a	a	r	f	f	a	a	a	a	a
a	a	a	f	r	f	f	a	a	a	a	a
a	a	a	r	r	f	f	a	a	a	a	a
a	a	a	r	f	f	f	a	a	a	a	a
a	f	f	r	f	f	f	a	a	a	a	a
a	f	f	r	f	u	f	a	a	a	a	a
h	h	h	h	h	h	h	h	h	h	h	h
f	f	r	u	u	u	u	a	a	a	a	a
f	f	r	f	u	u	a	a	a	a	a	a
f	f	f	r	f	f	a	a	a	a	a	a
f	f	f	f	r	f	a	a	a	a	a	a

a = agriculture u = developed
f = forest r = river
h = highways

Figure 2-32: Discrete or categorical data may be represented by codes in a raster data layer.

Table 2-1: Types of data represented by raster cell values.

Data Type	Description	Example
point ID	alpha-numeric ID of closest point	hospital
line ID	alpha-numeric ID of closest line	nearest road
contiguous region ID	alpha-numeric ID for dominant region	state
class code	alpha-numeric code for general class	vegetation type
table ID	numeric position in a table	row
physical analog	numeric value representing surface value	elevation
statistical value	numeric value from a statistical function	population density

lines, as the IDs of lines or points that occur closest to the cell center.

Point and line assignment to raster cells may be complicated when there are multiple features within a single cell. For example, when light poles are represented in a raster data layer, cell value assignment is straightforward when there is only one light in a cell (Figure 2-33, near A). When there are multiple poles in a single cell there is some ambiguity, or generalization in the assignment (Figure 2-33, near B). One common solution represents one feature from the group, and retains information on the attributes and characteristics of that feature. This entails some data loss. Another solution is to reduce the raster cell size so that there are no multiple features in a cell. This may result in impractically large data sets. More complex schemes may record multiple instances of features in a cell, but these then may slow access or otherwise decrease the utility that comes from the simple raster structure.

Similar problems may occur when there are multiple line segments within a raster cell, for example, when linear features such as roads are represented in a raster data set. When two or more roads meet, they will do so within a raster cell, and some set of attributes must be assigned (Figure 2-33 C). Since attributes are assigned by cells, some precedence must be established, with one line given priority over others.

Raster cell assignment also may be complicated when representing what we typically think of as discrete, uniform areas. Consider the area in Figure 2-34. We wish to represent this area with a raster data layer,

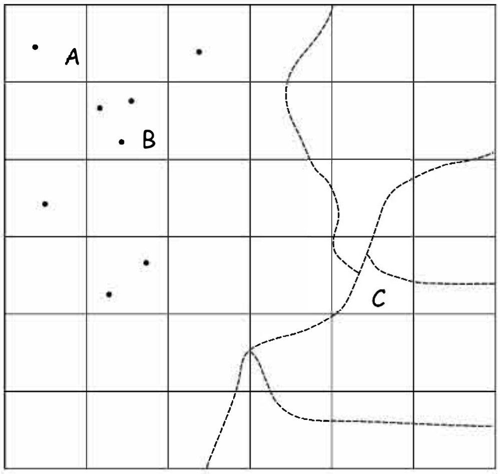


Figure 2-33: Raster cell assignment requires decisions when multiple objects occur in the same cell.

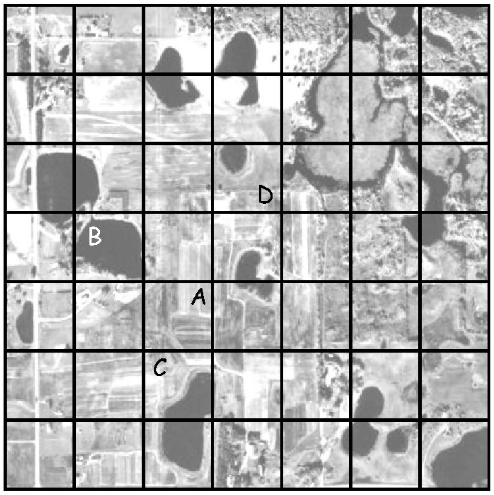


Figure 2-34: Raster cell assignment with mixed landscapes. Upland areas are lighter greys, water the darkest greys.

with cells assigned to one of two class codes, one each for land or water. Water bodies appear as darker areas in the image, and the raster grid is shown overlain. Cells may contain substantial areas of both land and water, and the proportion of each class may span from zero to 100 percent. Some cells are purely one class and the assignment is unambiguous; for example, the cell labelled A in the Figure 2-34 contains only land. Others are ambiguous, such as cell B (water) or D (land). Some are nearly equal in their proportion of land and water, as in cell C.

One common method to assign classes for mixed cells is called "winner-take-all". The cell is assigned the class of the largest area type. Cells A, C, and D would be assigned the land class, cell B the water class. Another option applies preference in cell assignment. If any of an "important" type is found, then the cell is assigned that value, regardless of the proportion. If we specify a preference for water, then cells B, C, and D in Figure 2-34 would be assigned the water type, and cell A the land type.

Regardless of the assignment method used, Figure 2-34 illustrates two phenomena when discrete objects are represented using a raster data model. First, some areas that are not the assigned class are included in some rasters cells. These "inclusions" are inevitable because cells must be assigned to a discrete class, the cell boundaries are rigidly assigned, and the class boundaries on the ground rarely line up with the cell boundaries. Some mixed cells occur in nearly all raster layers. The GIS user must acknowledge these inclusions, and consider their impact on the intended spatial analyses.

Second, differences in class assignment rules may substantially alter the data layer, as shown in our simple example. In more complex landscapes, there will be more potential cell types, which may increase the assignment sensitivity. Decreasing the raster cell size reduces the significance of classes in the assignment rule, but at the cost of increased data volumes.

The occurrence of more than one line or point within a raster cell can result in similar assignment problems. If two points occur, then which point ID is assigned? If two lines occur, then which line ID should be assigned? Some rule must be developed; for example, the point that falls nearest the center may be assigned, or the line with the longest segment within the raster cell. Similar to when area features are assigned to rasters, inclusions, and dependence on the class assignment rules affect the output.

Raster Features and Attribute Tables

Raster layers may also have associated attribute tables. This is most common when nominal data are represented, but may also be used with ordinal or interval/ratio data. Just as with topological vector data, features in the raster layer may be linked to rows in an attribute table, and these rows may describe the essential nonspatial characteristics of the features.

Figure 2-35a and b show data represented in a raster model. Figure 2-35a shows a raster data set that maintains a one-to-one relationship between raster cells and in the data table. An additional column, cell-ID,

must be added to uniquely identify each raster location. The corresponding attributes IDorg, class, and area are repeated for each cell. Note that the area values are the same for all cells and thus all rows in the table.

A one-to-one correspondence is rarely used with raster data sets because it often would require an unmanageably large size of attribute table. This small example results in 100 rows for the attribute table, but we often use raster data sets with billions of cells. If we insist on a one-to-one cell/attribute relationship, the table may become too large. Even simple processes such as sorting, searching, or subsetting records become prohibitively time consuming. Display and redraw rates become low, reducing the utility of these data, and decreasing the likelihood that GIS will be effectively applied.

To avoid these problems, a many-to-one relationship is usually allowed between the raster cells and the attribute table (Figure 2-35b). Many raster cells may refer to a single row in the attribute column. This substantially reduces the size of the attribute table for most data sets, although it does so at the cost of some spatial ambiguity. There may be multiple, noncontiguous patches for a specific type. For example, the upper left and lower right portion of the raster data set in Figure 2-35b are both of class 10. Both are recognized as distinct features in the vec-

a) Raster, one-to-one

A	A	A	A	B	B	B	B	B	B
A	A	A	A	B	B	B	B	B	B
A	A	A	A	B	B	B	B	B	B
A	A	A	B	B	B	B	B	B	B
A	A	A	C	C	B	B	B	B	B
C	C	C	C	C	D	D	D	D	D
C	C	C	C	C	D	D	D	D	D
C	C	C	C	C	D	D	D	D	D
C	C	C	C	C	D	D	D	E	E
C	C	C	C	C	D	D	E	E	E

attribute table
(cell 1 is upper-left corner)

cell-ID	IDorg	class	area
1	A	10	0.8
2	A	10	0.8
3	A	10	0.8
4	A	10	0.8
5	B	11	0.8
6	B	11	0.8
7	B	11	0.8
⋮	⋮	⋮	⋮
100	E	10	0.8

b) Raster, many-to-one

10	10	10	10	11	11	11	11	11	11
10	10	10	10	11	11	11	11	11	11
10	10	10	10	11	11	11	11	11	11
10	10	10	11	11	11	11	11	11	11
10	10	10	15	15	11	11	11	11	11
15	15	15	15	15	21	21	21	21	21
15	15	15	15	15	21	21	21	21	21
15	15	15	15	15	21	21	21	21	21
15	15	15	15	15	21	21	21	10	10
15	15	15	15	15	21	21	10	10	10

attribute table

class	area
10	18.4
11	24.0
15	21.6
21	13.6

Figure 2-35: Raster data models rarely maintain this one-to-one relationship between cells and attributes (a), because table access and performance usually suffer. A many-to-one relationship between cells and table rows is adopted more often (b).

tor and one-to-one raster representation, but are represented by the same attribute entry in the many-to-one raster representation. This reduces the size of the attribute table, but at the cost of reducing the flexibility of the attribute table. Many-to-one relationships effectively create multi-part areas. The data for the represented variable may be summarized by class; however, these classes may or may not be spatially contiguous.

An alternative is to maintain the one-to-one relationship, but to index all the raster cells in a contiguous group, thereby reducing the number of rows in the attribute table. This requires software to develop and maintain the indices, and to create them and reconstitute the indexing after spatial operations. These indexing schemes add overhead and increase data model complexity, thereby removing one of the advantages of raster data sets over vector data sets.

A Comparison of Raster and Vector Data Models

The question often arises, "Which are better, raster or vector data models?" The answer is neither and both. Neither of the two classes of data models is better in all conditions or for all data. Both have advantages and disadvantages relative to each other and to additional, more complex data models (Table 2-2). In some instances, it is preferable to maintain data in a raster model, and in others in a vector model. Most data may be represented in both, and may be converted among data models. As an example, land cover may be represented as a set of polygons in a vector data model or as a set of identifiers in each cell in a raster grid. The choice often depends on a number of factors, including the predominant type of data (discrete or continuous), the expected types of analyses, available storage, the main sources of input data, and the expertise of the human operators.

Table 2-2: A comparison of raster and vector data models.

Characteristic	Raster	Vector
data structure	usually simple	usually complex
storage requirements	larger for most data sets without compression	smaller for most data sets
coordinate conversion	may be slow due to data volumes, and require resampling	simple
analysis	easy for continuous data, simple for many layer combinations	preferred for network analyses, many other spatial operations more complex
spatial precision	floor set by cell size	limited only by positional measurements
accessibility	easy to modify or program, due to simple data structure	often complex
display and output	good for images, but discrete features may show "stairstep" edges	maplike, with continuous curves, poor for images

Raster data models exhibit several advantages relative to vector data models. First, raster data models are particularly suitable for representing themes or phenomena that change frequently in space. Each raster cell may contain a value different than its neighbors. Thus, trends as well as more rapid variability may be represented.

Raster data structures are generally simpler than vector data models, particularly when a fixed cell size is used. Most raster models store cells as sets of rows, with cells organized from left to right, and rows stored from top to bottom. This organization is quite easy to code in an array structure in most computer languages.

Raster data models also facilitate easy overlays, at least relative to vector models. Each raster cell in a layer occupies a given position corresponding to a given location on the Earth's surface. Data in different layers align cell-to-cell over this position. Thus, overlay involves locating the desired grid cell in each data layer and comparing the values found for the given cell location. This cell look-up is quite rapid in most raster data structures, so layer overlay is quite simple and rapid when using a raster data model.

Finally, raster data structures are the most practical method for storing, displaying, and manipulating digital image data, such as aerial photographs and satellite imagery. Digital image data are an important source of information when building, viewing, and analyzing spatial databases. Image display and analysis are based on raster operations to sharpen details on the image, specify the brightness, contrast, and colors for display, and to aid in the extraction of information.

Vector data models provide some advantages relative to raster data models. First, vector models often lead to more compact data storage, particularly for discrete objects. Large homogenous regions are recorded by the coordinate boundaries in a vector data model. These regions are recorded as a set of cells in a raster data model. The perimeter grows more slowly than the area for most feature shapes, so the amount of data required to represent an area increases much more rapidly with a raster data model. Vector data are much more compact than raster data for most themes and levels of spatial detail.

Vector data are a more natural means for representing networks and other connected linear features. Vector data by their nature store information on intersections (nodes) and the linkages between them (lines). Traffic volume, speed, timing, and other factors may be associated with lines and intersections to model many kinds of networks.

Vector data models are easily presented in a preferred map format. Humans are familiar with continuous line and rounded curve representations in hand- or machine-drawn maps, and vector-based maps show these curves. Raster data often show a "stair-step" edge for curved boundaries, particularly when the cell resolution is large relative to the resolution at which the raster is displayed. Vector data may be plotted with more visually appealing continuous lines and rounded edges.

Vector data models facilitate the calculation and storage of topological information. Topological information aids in performing adjacency, connectivity, and other analyses in an efficient manner. Topological information also allows some forms of automated error and ambiguity detection, leading to improved data quality.

Conversion Between Raster and Vector Models

Spatial data may be converted between raster and vector data models. Vector-to-raster conversion involves assigning a cell value for each position occupied by vector features. Vector point features are typically assumed to have no dimension. Points in a raster data set must be represented by a value in a raster cell, so points have at least the dimension of the raster cell after conversion from vector-to-raster models. Points are usually assigned to the cell containing the point

coordinate. The cell in which the point resides is given a number or other code identifying the point feature occurring at the cell location. If the cell size is too large, two or more vector points may fall in the same cell, and either an ambiguous cell identifier assigned, or a more complex numbering and assignment scheme implemented. Typically, a cell size is chosen such that the diagonal cell dimension is smaller than the distance between the two closest point features.

Vector line features in a data layer may also be converted to a raster data model. Raster cells may be coded using different criteria. One simple method assigns a value to a cell if a vector line intersects with any part of the cell (Figure 2-36a, left). This ensures the maintenance of connected lines in the raster form of the data. This assignment rule often leads to wider than appropriate lines because several adjacent cells may be assigned as part of the line, particularly when the line meanders near cell edges. Other assignment rules may be applied, for example, assigning a cell as occupied by a line only when the cell center is near a vector line segment (Figure 2-36a, right). "Near" may be defined as some sub-cell distance, for instance, 1/3 the cell width. Lines passing through the corner of a cell will not be recorded as in the cell. This may lead to thinner linear features in the raster data set, but often at the cost of line discontinuities.

The output from vector-to-raster conversion depends on the algorithm used, even though you use the same input. This brings up an important point to remember when applying any spatial operation. The output

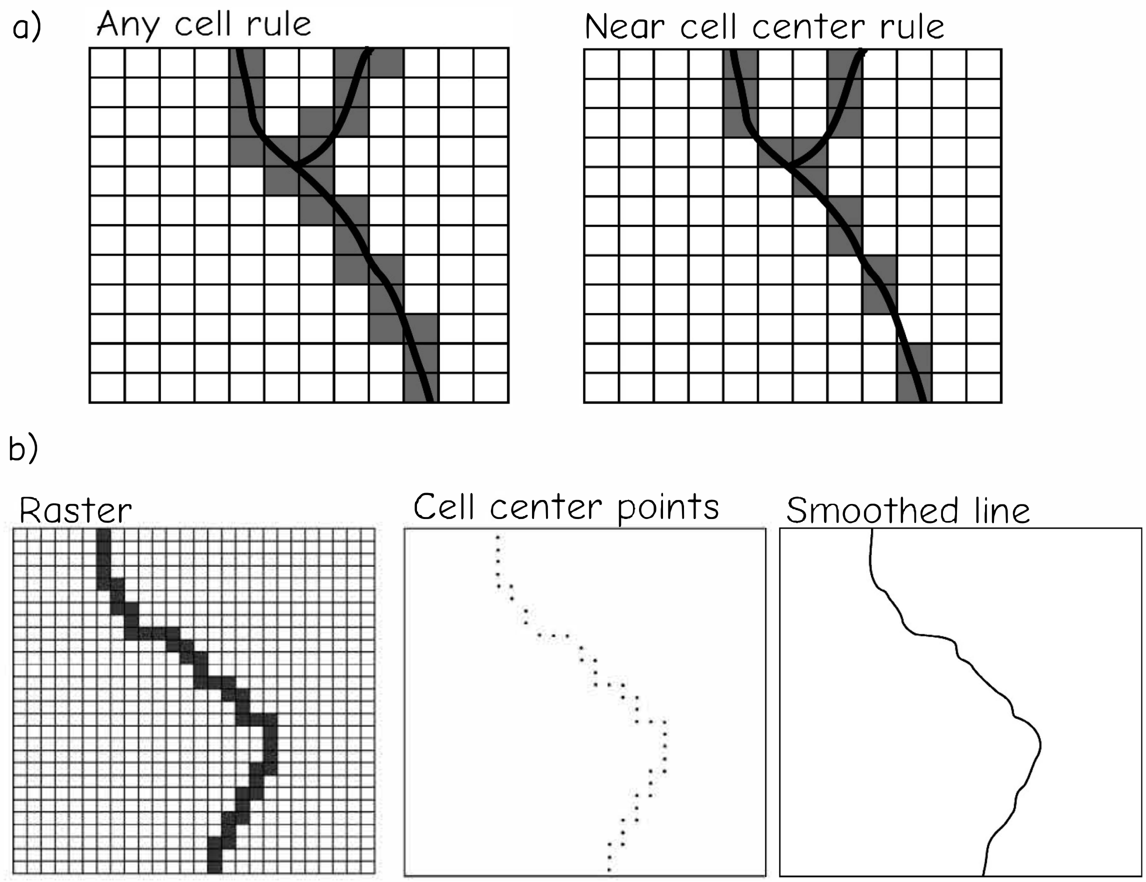


Figure 2-36: Vector-to-raster conversion (a) and raster-to-vector conversion (b). In a, cells are assigned in a raster if they intersect with a converted vector. The left and right panels show how two assignment rules result in different raster coding near lines. Panels in b show how raster data may be converted to vector formats, and may involve line smoothing or other operations to remove the "stair-step" effect.

often depends in subtle ways on the spatial operation. What appear to be quite small differences in the algorithm or key defining parameters may lead to quite different results. The ease of spatial manipulation in a GIS provides a powerful and often easy-to-use set of tools. The GIS user should bear in mind that these tools can be more efficient at producing errors as well as more efficient at providing correct results. Until sufficient experience is obtained with a suite of algorithms, in this case vector-to-raster conversion, small, controlled tests should be performed to verify the accuracy of a given method or set of constraining parameters.

Up to this point we have covered vector-to-raster data conversion. Data may also be converted in the opposite direction, from raster to vector data. Point, line, or area features represented by raster cells are converted to corresponding vector data coordinates and structures. Point features are represented as single raster cells. Each vector point feature is usually assigned the coordinate of the corresponding cell center.

Linear features represented in a raster environment may be converted to vector lines. Conversion to vector lines typically involves identifying the continuous connected set of grid cells that form the line. Cell centers are typically taken as the locations of vertices along the line (Figure 2-36b). Lines may then be "smoothed" using a mathematical algorithm to remove the "stair-step" effect.

Other Data Models

Triangulated Irregular Networks

A *triangulated irregular network* (TIN) is a data model commonly used to represent terrain heights. Typically, the x, y, and z locations for measured points are entered into the TIN data model. These points are connected such that the smallest triangle possible spans any three adjacent points. The TIN forms a connected network of triangles (Figure 2-37). *Delaunay triangles* are created such that the line crossings are avoided. Triangulation identifies the *convergent circle* for a set of three points, defined as a circle passing through all three points. A triangle is drawn only if the corresponding convergent circle contains no other sampling points. Each triangle defines a facet of uniform slope and aspect over the triangle.

The TIN model typically uses some form of indexing to connect neighboring points. Each edge of a triangle connects to two points, which in turn each connect to other edges. These connections continue recursively until the entire network is spanned.

While the TIN model may be more complex than simple raster models, it is often more efficient for storing terrain data in areas with variable relief. Relatively few points are required to represent large, flat, or smoothly continuous areas. Many more points are desirable for rugged terrain. Surveyors often collect more samples per unit area where the terrain is highly variable. A TIN easily accommodates these differences in sampling density, resulting in more, smaller triangles in the densely sampled area.

Object Data Models

The *object data model* is an alternative for structuring spatial data. A main goal is to raise the level of abstraction so that the data objects may be conceptualized and addressed in a more natural way. Objects are often geographic features, with spatial and attribute data associated with the object, e.g., a city object may include information on the city boundary, streets, building locations, waterways, or other features in organized

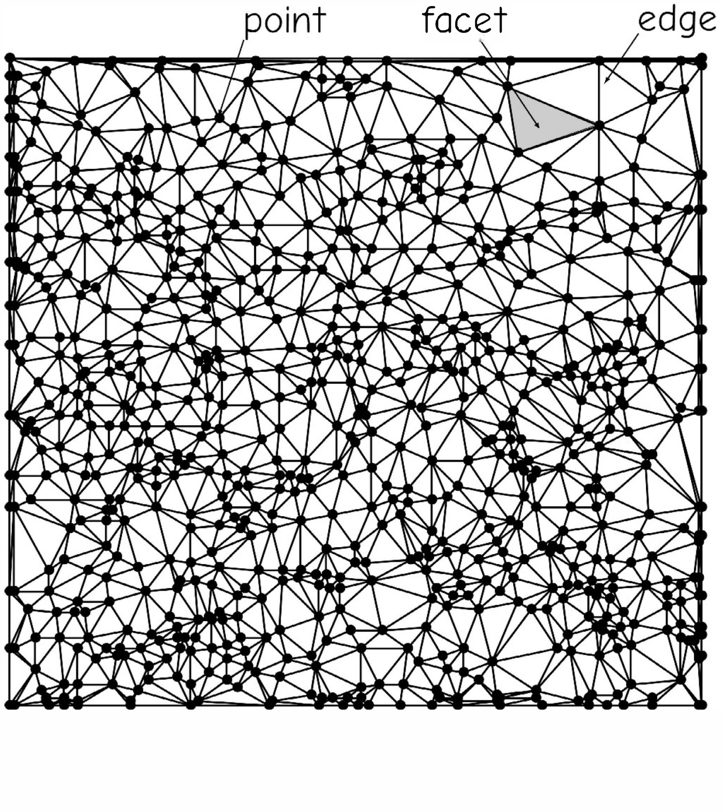
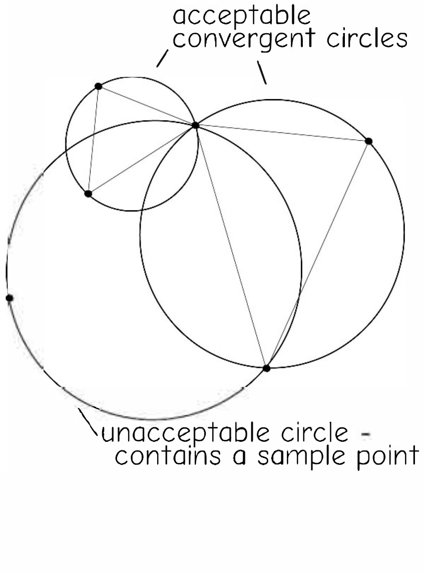


Figure 2-37: A TIN data model defines a set of adjacent triangles over a sample space (left). Sample points, facets, and edges are components of TIN data models. Triangles are placed by convergent circles. These intersect the vertices of a triangle and contain no other possible vertices (below).



data structures. Vector topology could be included, incorporated within the single object. Object model approaches have been adopted by at least one major vendor of GIS software and are applied in a number of fields.

Object models for spatial data often follow a *logical model*, a user's view of the real objects we portray with a GIS (Figure 2-38). This model includes all the "things" of interest, and the relationships among them. Things, or objects, might include power poles, transformers, powerlines, meters, and customer buildings in a city, and relationships among them would include a transformer on a pole, lines between poles, and meters at points along the lines. The logical model is often represented as a box-and-line diagram.

Most object models define the properties of each object, and the relationships among objects. Pipe objects may have a diameter, material type, and be connected to valves and tanks. The pipes may be represented by lines and the valves by points, but these vector elements are enhanced in the object model because the specific pipe and pipe properties may be linked to the specific valve attached to a given location. Object models can have *inheritance*, automatically transfering properties within classes of objects. We may create a generic valve object, with a maximum pressure rating, cost, and material type. We may create valve subclasses within this class, e.g., emergency cut-off valves, primary control valves, or shunt valves. These subclasses will inherit all the property variables from a generic valve in that each has a cost, maximum pressure, and material, but each subclass may also have additional, unique properties.

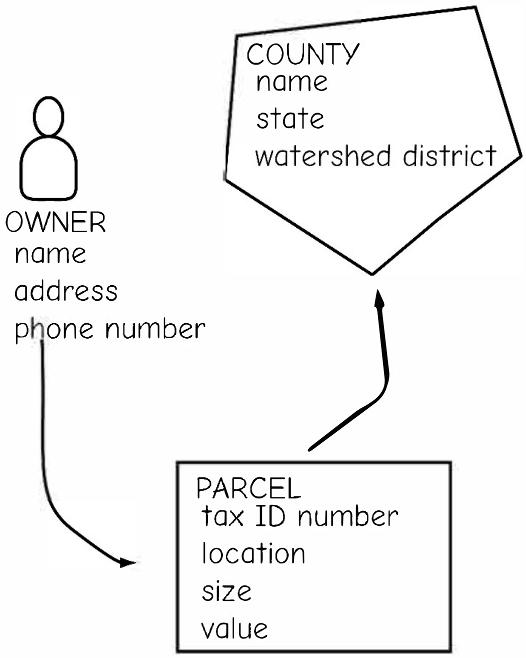


Figure 2-38: Objects in a GIS database may be conceptualized in a diagram, or logical model of how they are related. Here, three types of objects are represented, with owners associated with parcels, and parcels associated with counties.

Figure 2-39 shows an example of an object data model for hydrologic basins and related stream features. The top frame shows features; in this example basins, sub-basins, a stream network, and features on the stream network such as sampling stations. The bottom panel shows the feature types, attributes, and properties in the object model. Note that there are both object properties and topological relationships represented, and that multiple feature types may be represented in the object model. The object data model has both advantages and disadvantages when compared to traditional data models. Some geographic entities may be naturally and easily identified as discrete units for particular problems, and so may be naturally amenable to an object oriented approach. Some proponents claim object models are more easily implemented across a wider range of database software, particularly for complex models. However, object data models are less useful for representing continuously varying features, such as elevation. In addition, for many problems, object definition and indexing may be quite complex. Software developers have had difficulty developing generic tools that may quickly implement object models, so there is an added level of specialized training required. Finally, we note that there is no widely accepted, formal definition of what constitutes an object data model.

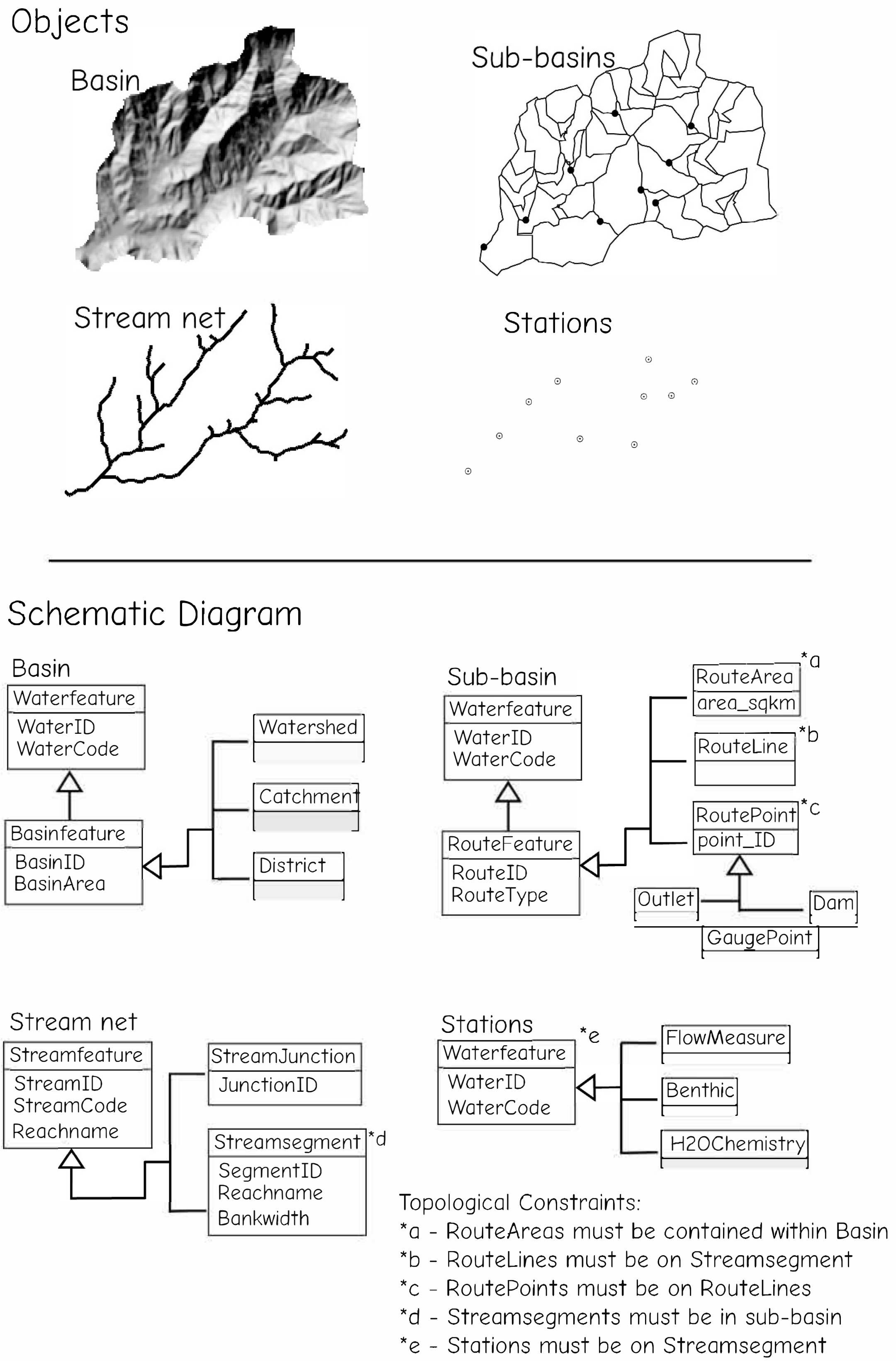


Figure 2-39: Object-oriented data models allow us to encapsulate complex objects that may be a combination of many different features and feature types, while explicitly identifying the embedded complexity in a standard way. Constraints such as topological relationships across objects may also be represented.

Three-Dimensional Data Models

GIS in built environments are increasingly integrating three-dimensional (3D) information such as building heights, roof shapes, and other height-related characteristics (Figure 2-40). This is in part to support analysis, and in part to generate visualizations from at or near ground-level, for example, building appearance from a nearby road. Improved data capture allows for the rapid development of 3D data that must be integrated into an appropriate data model.

Several 3D data models have been proposed, with "vector-like" models more commonly applied than "raster-like" models. The latter employ the concept of voxels, or volume elements, essentially cubes of a fixed dimension. Flat or baseline areas have zero voxels stacked over the surface, with a requisite number of voxels "stacked" at each raster cell to represent height. While easy to access and simple for all the reasons a raster set is, they also suffer from the same drawbacks, particularly the trade-off between precision, or voxel size, and data volumes.

Many alternative "vector-like" 3D models have been proposed, many defining a body element, in addition to the standard point, line, and polygon elements. Points are used to create lines, lines for polygons, and polygons to create bodies. Much like 2D vectors, 3D vectors add indexing schemes, with the added complication of a z coordinate to any element above the base plane of the data layer.

One 3D model, developed by Bentley Systems, is called a *reality mesh* (Figure 2-41). It combines a three-dimensional triangulated irregular network with an image-based texture surface to create realistic representations of 3D features. Complex 3D surfaces can be efficiently represented, demanding less storage space, while properties like material or color can be associated

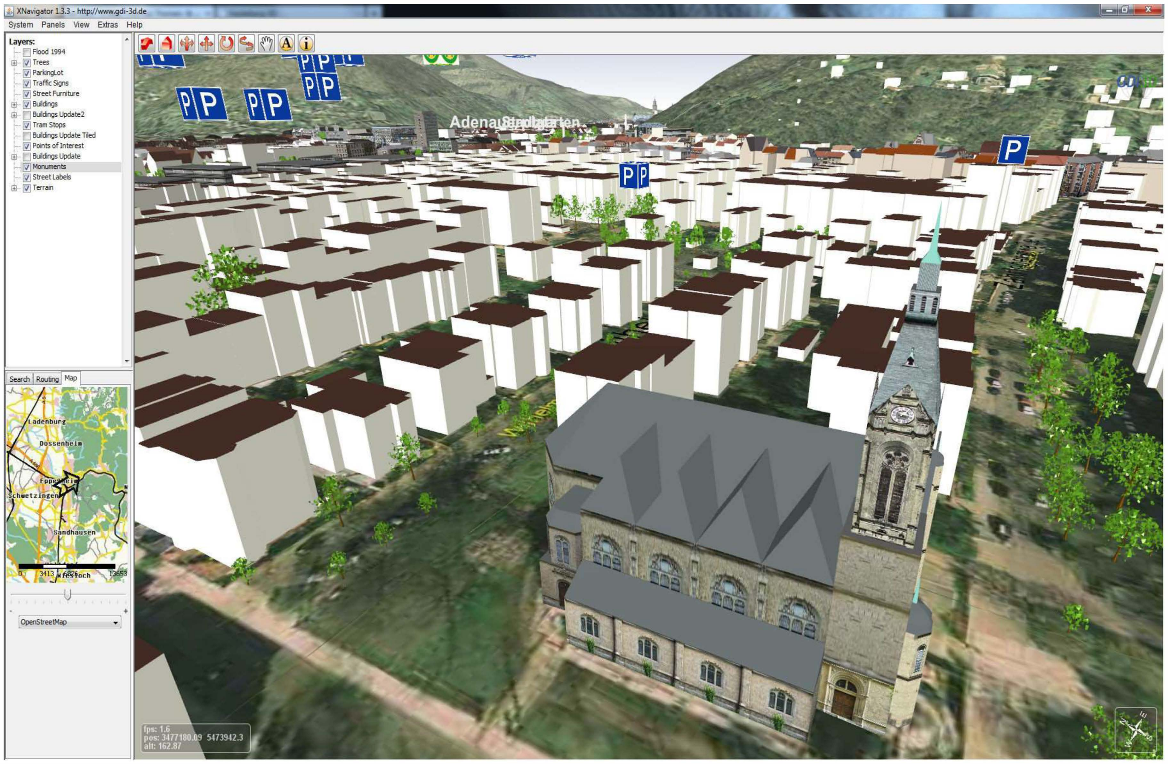


Figure 2-40: An example of 3D spatial data displayed for a region in Germany. The third dimension aids in many planning and assessment functions, and is becoming common in managing the built environment. (courtesy City of Heidelberg and the University of Heidelberg).

with each of the triangular faces. A compressed raster "texture surface" may be wrapped around the 3D TIN, yielding substantial detail. The images may be compressed, using methods described later in this chapter, to reduce data volumes.

While vector 3D models are becoming common, no one model form or standard has been widely adopted. Three dimensional GIS products have become quite mature, for example, the 3D spatial analyst and 3D CityEngine products from ESRI are full-featured, stable, productive tools, supporting start to finish workflows, from 2D to 3D data conversion, 3D spatial data ingestion, processing and organization, and output to project, video, and interactive Web services. Three-dimensional GIS processing tools are available from other vendors, and as a plug-in for QGIS.

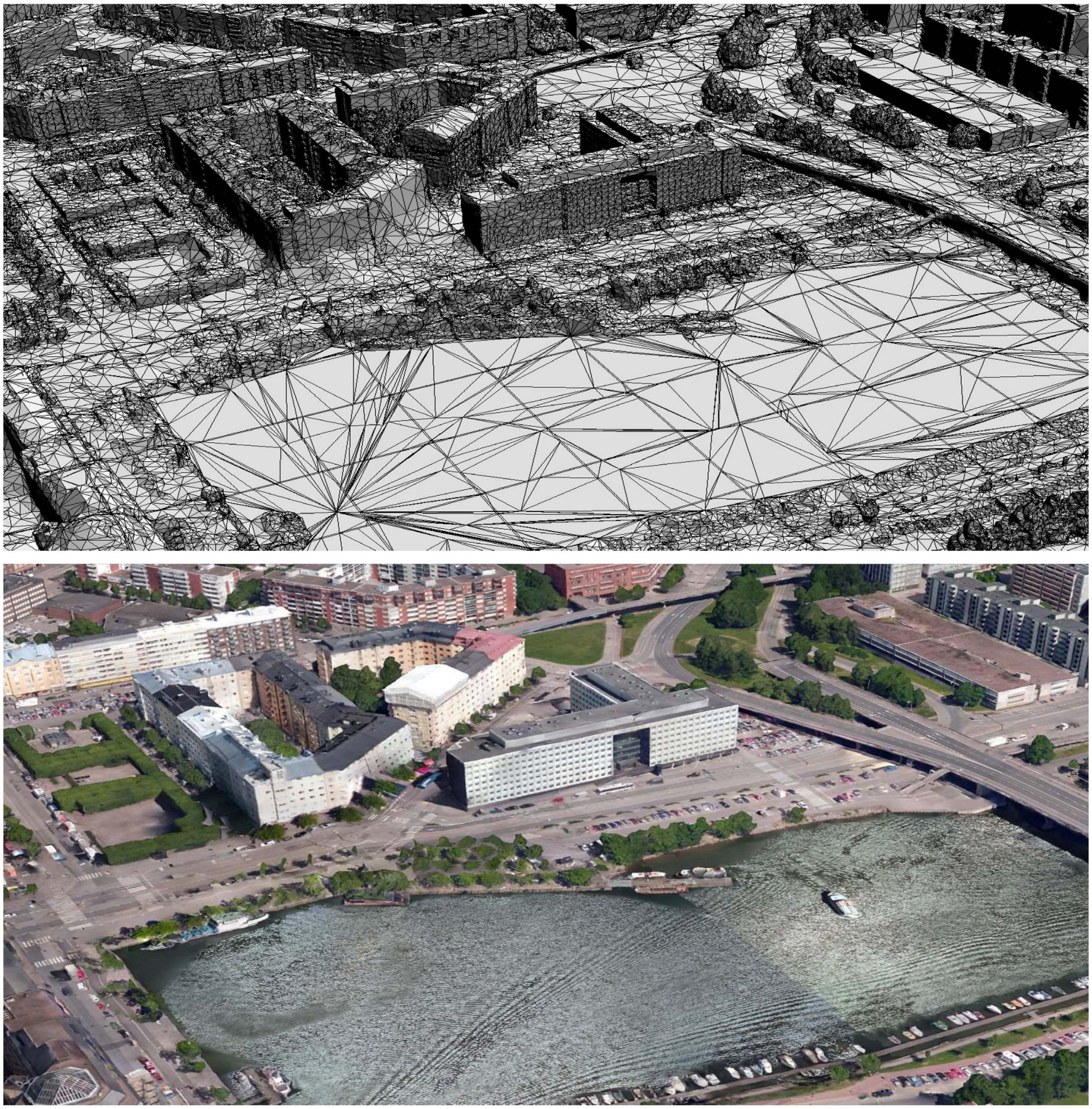


Figure 2-41: An example of a reality mesh as a 3-D model. Surface geometry is recorded in a 3-D triangulated irregular network, while a corresponding "texture" surface is projected on to corresponding facets (courtesy Bentley Systems).

Multiple Models

Digital data may often be represented using any one of several data models. The analyst must choose which representation to use. Digital elevation data are perhaps the best example of the use of multiple data models to represent the same theme (Figure 2-42). Digital representations of terrain height have a long history and widespread use in GIS. Elevation data and derived surfaces such as slope and aspect are important in hydrology, transportation, ecology, urban and regional planning, utility routing, and a number of other activities that are analyzed using GIS. Because of this widespread importance, digital elevation data are commonly represented in a number of data models.

Raster grids, TINs, and vector contours are the most common data structures used to organize and store digital elevation data. Raster and TIN data are often called *digital elevation models* (DEMs) or *digital terrain models* (DTMs), and are commonly used in terrain analysis. Contour lines are most often used as a form of input, or as a familiar form of output. Historically, hypsography (terrain height) was depicted on maps as contour lines (Figure 2-42). Contours represent lines of equal elevation, typically spaced at fixed elevation intervals across the mapped areas. Because many important analyses are more difficult using contour lines, most digital elevation data are stored using raster or TIN models.

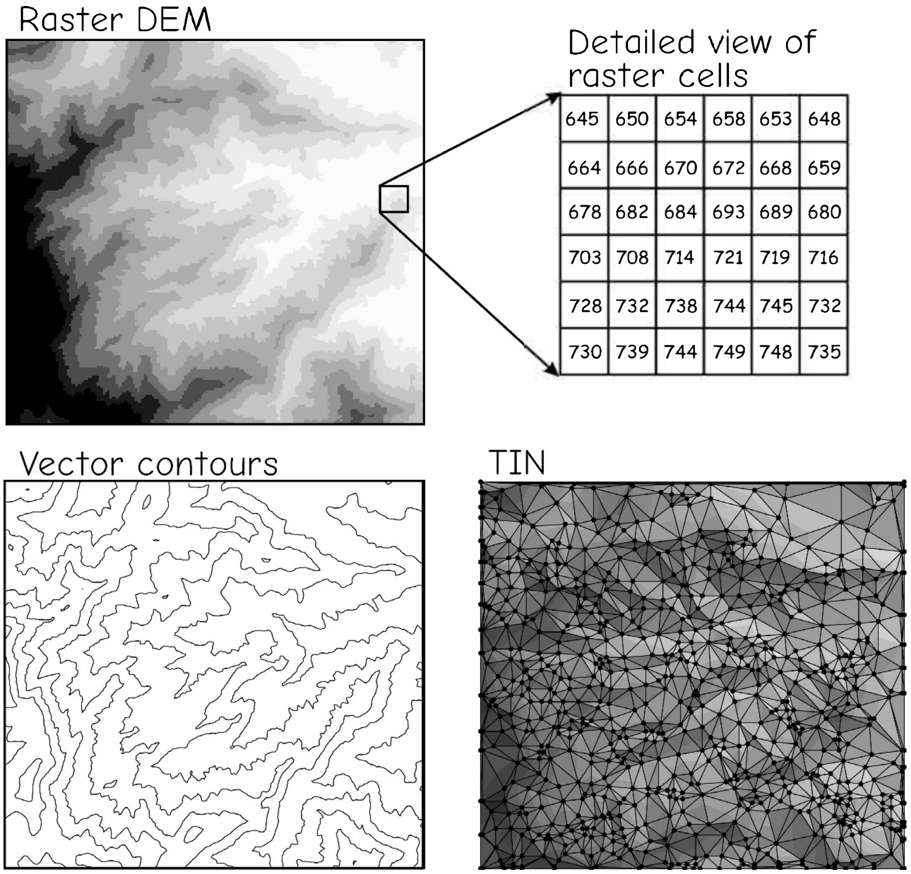


Figure 2-42: Data may often be represented in several data models. Digital elevation data are commonly represented in raster (DEM), vector (contours), and TIN data models.

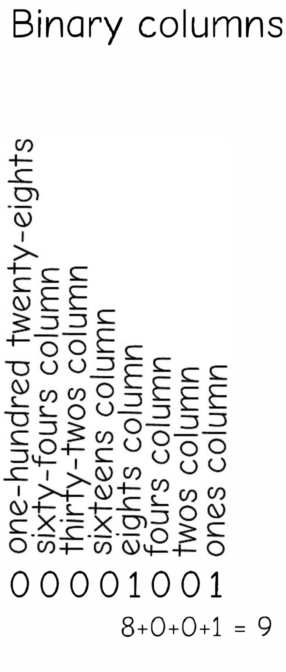
Data and File Structures

Binary and ASCII Numbers

No matter which spatial data model is used, the concepts must be translated into a set of numbers stored on a computer. All information stored on a computer in a digital format may be represented as a series of 0's and 1's. These data are said to be stored in a *binary* format, because each digit may contain one of two values, 0 or 1. Binary numbers are in a base of 2, so each successive column of a number represents a power of two.

We use a similar column convention in our familiar ten-based (decimal) numbering system. As an example, consider the number 47, which we represent using two columns. The seven in the first column indicates there are seven units of one. The four in the tens column indicates there are four units of ten. Each higher column represents a higher power of ten. The first column represents one ($10^0=1$), the next column represents tens ($10^1=10$), the next column represents hundreds ($10^2=100$), and upward for successive powers of ten. We add up the values represented in the columns to decipher the number.

Binary numbers are also formed by representing values in columns. In a binary system each column represents a successively higher power of two (Figure 2-43). The first (rightmost) column represents 1 ($2^0 = 1$), the second column (from right) represents twos ($2^1 = 2$), the third (from right) represents fours ($2^2 = 4$), then eight ($2^3 = 8$), sixteen ($2^4 = 16$), and upward for successive powers of two. Thus, the binary number 1001 represents the decimal number 9: a one from the rightmost column, and eight from the fourth column (Figure 2-43, left).



Equivalent numbers

binary	decimal
00000001	1
00000010	2
00000011	3
00000100	4
00000101	5
00000110	6
00000111	7
00001000	8
00001001	9
00001010	10
....	....

Figure 2-43: Binary representation of decimal numbers.

Each digit or column in a binary number is called a *bit*, and eight columns, or bits, are called a *byte*. A byte is a common unit for defining data types and numbers, for example, a data file may be referred to as containing 4-byte integer numbers. This means each number is represented by 4 bytes of binary data (or 8 x 4 = 32 bits).

Several bytes are required when representing larger numbers. For example, one byte may be used to represent 256 different values. When a byte is used for nonnegative integer numbers, then only values from 0 to 255 may be recorded. This will work when all values are below 255, but consider an elevation data layer with values greater than 255. If the data are not rescaled, then more than one byte of storage is required for each value. Two bytes will store up to 65,536 different numbers. Terrestrial elevations measured in feet or meters are all below this value, so two bytes of data are often used to store elevation data. Real numbers such as 12.19 or 865.3 typically require more bytes, and are effectively split, that is, two bytes for the whole part of the real number, and four bytes for the fractional portion.

Binary numbers are often used to represent codes. Spatial and attribute data may then be represented as text or as standard codes. This is particularly common when raster or vector data are converted for export or import among different GIS software systems. For example, ArcGIS, a widely used GIS, produces several export formats that are in text or binary formats. Idrisi, another popular GIS, supports binary and alphanumeric raster formats.

One of the most common number coding schemes uses ASCII designators. ASCII stands for the American Standard Code for Information Interchange. ASCII is a standardized, widespread data format that uses seven bits, or the numbers 0 through 126, to represent text and other characters. An extended ASCII, or ANSI (American National Standards Institute) scheme, uses these same codes, plus an extra binary bit to represent numbers between 127 and 255. These codes are then used in many programs, including GIS, particularly for data export or exchange.

ASCII codes allow us to easily and uniformly represent alphanumeric characters such as letters, punctuation, other characters, and numbers. ASCII converts binary numbers to alphanumeric characters through an index. Each alphanumeric character corresponds to a specific number between 0 and 255, which allows any sequence of characters to be represented by a number. One byte is required to represent each character in extended ASCII coding, so ASCII data sets are typically much larger than binary data sets. Geographic data in a GIS may use a combination of binary and ASCII data stored in files. Binary data are typically used for coordinate information, and ASCII or other codes may be used for attribute data.

Pointers and Indexes

Data files may be linked by file *pointers*, *indexes*, or other structures. A pointer is an address or index that connects one file location to another. Pointers are a common way to organize information within and across multiple files. Figure 2-44 depicts an example of the use of pointers to organize spatial data. In this figure, the polygon is composed of a set of lines. Pointers are used to link the set of lines that form each polygon. There is a pointer from each line to the next line, forming a chain that defines the polygon boundary.

Pointers help by organizing data in such a way as to improve access speed. Unorganized data would require time-consuming searches each time a polygon boundary was to be identified. Pointers also allow efficient use of storage space. In our example, each line segment is stored only once. Several polygons may point to the line segment, as it is typically much more space efficient to add pointers than to duplicate the line segment.

Shapefiles are a common vector spatial data format that uses an index to link files. Shapefiles were originally developed by ESRI as a way to store point, line, and polygon features, although they have since been

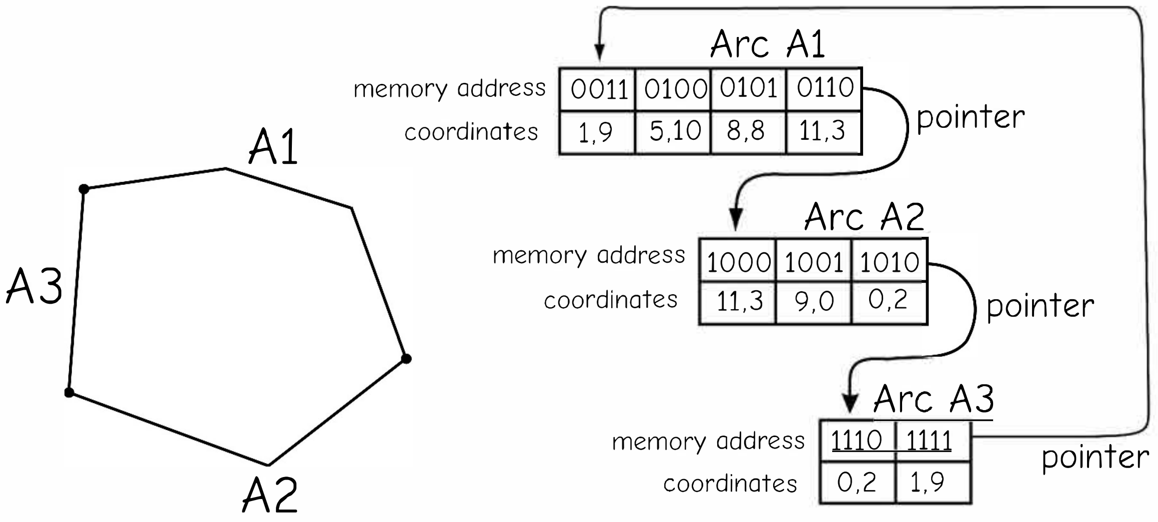


Figure 2-44: Pointers are used to organize vector data. Pointers reduce redundant storage and increase speed of access.

adopted as a common format for data interchange and analysis. Shapefiles are supported by most other GIS softwares that process vector data.

Shapefiles represent layers with a cluster of files. Each file has the same base name but a different filename extension, indicated by a suffix, for example, the ".shp" in the filename "boundary.shp." A transportation data layer stored in shapefile format might have the base name of roads, with different suffixes for different files:

roads.shp

roads.shx

roads.dbf

roads.prj

etc.

The first three files above are all required to represent a vector data layer using shapefiles. These files are connected using indices, numbers that identify connections and groupings for various components. The .shp files contain the coordinates that represent each road, organized by line segments. There is general information for each segment, and then a list of coordinates and other data for the segment. This is followed by general information for the next segment, and another list. The roads.shx file contains indices that point to the segment records in the .shp files, based on these identifiers. This speeds access, avoiding a search in the shp file each time it links segments in a road. The roads.dbf file also uses an index to point to the combined roads in the shp and shx files. A group of segments may be used to form a line, and associated with a set of attributes stored in a dbf file.

Because pointers and indices are key elements in organizing the spatial data, altering them directly will usually cause problems. Typically, these indices are created by the software during processing, and updated as needed when data are added, modified, or analyzed. Pointers may be visible, for example, the OID columns in the .dbf tables used with shapefiles, but manually changing the values will often ruin the data layer. You should know the identity and use of pointers in your data sets, so that you don't change them inadvertently.

Pointers, indexing, and multifile layers are not limited to vector data. Many raster formats store a majority of the cell data in one file, and additional, linked information in an associated file. You must be careful when transferring a data layer to include all the associated files. For example, copying

the roads.shp and roads.dbf files to a new location does not copy a usable data layer. The software expects a .shx file; an incomplete file set is often useless.

Two more complex structures are common, ESRI Geodatabases and the Open Geospatial Consortium (OGC) GeoPackage. These are proprietary and open standards, respectively, for storing both vector and raster data and topologies in an integrated fashion.

Data Compression

We often compress spatial data files because they are large. Data compression reduces file size while maintaining the information contained in the file. Compression algorithms may be *lossless*, in that all information is maintained during compression, or *lossy*, in that some information is lost. A lossless compression algorithm will produce an exact copy when a file is compressed and decompressed. A lossy algorithm will alter the data on a compression- decompression round trip. Lossy algorithms are most often used with image data, where substantial degradation still leaves a useful image, and are uncommonly applied to thematic spatial data, where any data degradation is typically not tolerated.

Data compression is most often applied to discrete raster data, for example, when representing polygon or area information in a raster GIS. There are redundant data elements in raster representations of large homogenous areas. Each raster cell within a homogenous area will have the same code as most or all of the adjacent cells. Data compression algorithms remove much of this redundancy.

Run-length coding is a common data compression method. This compression technique is based on recording sequential runs of raster cell values. Each run is recorded as the value found in the set of adjacent cells and the run length, or number of cells with the same value. Seven sequential cells of type A might be listed as A7 instead of AAAAAAA. Thus, seven cells would be represented by two characters. Consider the data recorded in Figure 2-45, where each line of raster cells is represented by a set of run-length codes. In general, run-length coding reduces data volume, as shown for the top three rows in Figure 2-45. Note that in some instances, run-length coding increases the data volume, most often when there are no long runs. This occurs in the last line of Figure 2-45, where frequent changes in adjacent cell values result in many short runs. However, for most thematic data sets containing area information, run-length coding substantially reduces the size of raster data sets.

There is also some data access cost in run-length coding. Standard raster data access involves simply counting the number of cells across a row to locate a given cell. To locate a cell in run-length coding we must

Raster

9	9	6	6	6	6	6	7
6	6	6	6	6	6	6	6
9	9	6	6	6	6	7	7
9	8	9	6	6	7	7	5

Run-length codes

2:9, 5:6, 1:7

8:6

2:9, 4:6, 2:7

1:9, 1:8, 1:9, 2:6, 2:7, 1:5

Figure 2-45: Run-length coding is a common and relatively simple method for compressing raster data. The left number in the run-length pair is the number of cells in the run, and the right is the cell value. Thus, the 2:9 listed at the start of the first line indicates a run of length two for the cell value 9.

sum along the run-length codes to identify a cell position. This is typically a minor additional cost, but in some applications the trade-off between speed and data volume may be objectionable.

Quad tree representations are another raster compression method. Quad trees are similar to run-length codings in that they are most often used to compress raster data sets when representing area features. Quad trees may be thought of as a raster data structure with a variable spatial resolution. Raster cell sizes are combined and adjusted within the data layer to fit into each specific area feature (Figure 2-46). Large raster cells that fit entirely into one uniform area are assigned the value corresponding to that area; for example, the three largest cells in Figure 2-46 are all assigned the value a. Successively smaller cells are then fit, halving the cell dimension at each iteration, again fitting the largest cell that will fit in each uniform area. This is illustrated in the top-left corner of Figure 2-46. Successively smaller cells are defined by splitting "mixed cells" into four quadrants, and assigning the values a or b to uniform areas. This is repeated down to the smallest cell size that is needed to represent uniform areas at the required detail.

The varying cell size in a quad tree representation requires more sophisticated

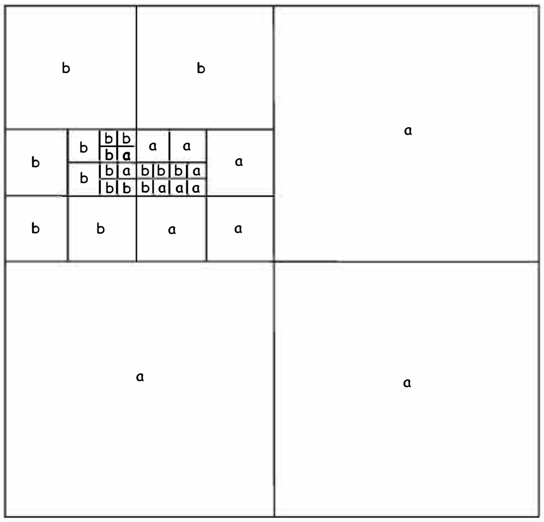


Figure 2-46: Quad tree compression.

indexing than simple raster data sets. Pointers are used to link data elements in a tree-like structure, hence the name quad trees. There are many ways to structure the data pointers, from large to small, or by dividing quandrants, and these methods are beyond the scope of an introductory text. Further information on the structure of quad trees may be found in the references at the end of this chapter.

There are many other data compression methods that are commonly applied. LZW is a lossless compression method commonly applied to image and raster data sets, particularly GIF images and TIFF formats. JPEG and wavelet compression algorithms are often applied to reduce the size of spatial data, particularly image or other data, although as implemented these are lossy algorithms. Generic bit- and byte-level compression methods may be applied to any files for compression or communications. There is usually some cost in time to the compression and decompression.

Raster Pyramids

We sometimes intentionally increase the size of our raster data sets without increasing the resolution in a process known as *pyramiding*. We create pyramids to increase display speeds when viewed at small scales ("zoomed out"). Long redraw times often hinder use of large data sets, particularly when panning frequently. When displayed at very small scales, the cell size of a data set may be smaller than the resolution of the computer screen. A raster data set 1,000,000 pixels across has 1,000 times the data that can be displayed on a monitor with 1,000 pixel horizontal resolution. However, display software must wade through all 1,000,000 data elements in a row to pick the 1 cell in 1,000 to display. While clever software can help, there are limits to how much we can speed up the redraws.

Pyramiding in effect saves subsampled copies of the cells at various resolutions. In our example above, pyramids may do the equivalent of saving every two, every four,

every 10, every 30, and every 100 cells, all within the same raster data set. The software then compares the display scale to the dimensions of the data set, and chooses the most appropriate cell resolution to display. Redraws are much faster, and transparent to the user.

Note that we say pyramids "in effect" save copies of cells at various resolutions. This is the simplest method, but often not the most efficient for space or speed of access. Sophisticated indexing may be used to point to the cells at the appropriate resolutions.

Note that pyramiding comes at a cost, both in the size and complexity of the raster data set. Indexing schemes complicate the simple raster data structure, and the software must be able to navigate the indexing scheme. Already large raster data sets may be inflated from a few percent to several times, although in practice it is typically less than a doubling of size.

Common File Formats

A few file formats are commonly used to store and transfer spatial data. Some of these file structures arose from distribution formats adopted by governmental agencies, others were specified by software vendors, and some have been devised by standards-making bodies. Some knowledge of the types and properties of these file formats is helpful to the GIS practitioner.

Common geographic data formats may be placed into three large classes: raster, vector, and attribute. Raster formats may be further split into single-band and multi-band file types. Multi-band raster data sets are most often used to store and distribute image data, while single-band raster data sets are used to store both single-band images and nonimage spatial data. Table 2-3 summarizes some of the most common spatial data formats.

Most GIS softwares provide some utility for data import and export from standard formats

The Geospatial Data Abstraction Library (GDAL) provides a utility to translate among many common vector and raster file formats. This free utility is flexible, and often can be used to extend the reach of commercial packages, by first using GDAL to convert files from unsupported to supported types, and then importing these into the target software.

Summary

In this chapter we have described the main ways of conceptualizing spatial entities, and of representing these entities as spatial features in a computer. We commonly employ two conceptualizations, also called spatial data models: a raster data model and a vector data model. Both models use a combination of coordinates, defined in a Cartesian or spherical system, and attributes, to represent our spatial features. Features are usually segregated by thematic type in layers.

Vector data models describe the world as a set of point, line, and area features. Attributes may be associated with each feature. A vector data model splits that world into discrete features, and often supports topological relationships. Vector models are most often used to represent features that are considered discrete, and are compatible with vector maps, a common output form.

Raster data models are based on grid cells and represent the world as a "checkerboard," with uniform values within each cell. A raster data model is a natural choice for representing features that vary continuously across space, such as temperature or precipitation. Data may be converted between raster and vector data models.

We use data structures and computer codes to represent our conceptualizations in more abstract, but computer-compatible forms. These structures may be optimized to reduce storage space and increase access speed, or to enhance processing based on the nature of our spatial data.

Table 2-3: Common formats for spatial data.

Type and source	Extension	Characteristics (R=Raster, V=Vector, A=Attribute, I=Image)
Comma Separated Value	.csv	Common ASCII text format used to distribute attribute and often vector information (A, V).
DXF, AutoDesk	.dxf	Drawing exchange file, an ASCII or binary file for exchanging spatial data (V).
DWG, AutoDesk	.dwg	Native binary file used by AutoDesk to store geographic data and drawings in AutoCAD (V).
Geodatabase, ESRI	.gdb, .mdb	ESRI container for many data types (R, V, A, I).
GeoJSON, open standard	.json, .geojson	Open standard for representing and displaying simple geographic features (V, A).
GeoPackage, open standard	.gpkg	Open standard for representing vector and raster data, compatible with SQlite (R, V, A).
GeoTIFF, open standard	.TIF, .TIFF	An extension for georeferencing Aldus-Adobe public domain TIFF format (R).
GPX, open standard	.gpx	A specification based on XML for basic GNSS data (V).
Imagine, ERDAS	.img	Multiband capable image format (R).
Interchange, ESRI	.e00	ASCII text file for vector and identifying attribute data (V).
Keyhole Markup Language, Google	.KML	XML extension for displaying and annotating features and images (V, I, A).
LAS, ASPRS	.LAS	Laser point cloud data storage (V).
Shapefile, ESRI	.shp, .shx, .dbf, .prj, and others	Three or more binary files that include the vector coordinate, attribute, and other information (V).
TIGER, U.S. Census	tgrxxyyy, stfzz	Set of files by U.S. census areas, xx is a state code, yyy an area code, zz numbers for various file types (V, A).
MIF/MID, MapInfo	.mif, .mid	Map Interchange File, vector and raster data transport from MapInfo (V,R).
NetCDF, OGC	.cdf	Machine-independent data formats for scientific data arrays (R, A, I).
NLAPS, NASA	various in a directory	Image data from various Landsat satellites, in a specified directory structure (I, R).
SDTS, U.S. Government	none	Spatial Data Transfer Standard, specifies the spatial objects, attributes, reference system (R, V, A).

Suggested Reading

Arctur, D., Zeiler, M. (2004). *Designing Geodatabases: Case Studies in GIS Data Modeling*. Redlands: ESRI Press.

Batcheller, J.K., Gittings, B.M., Dowers, S. (2007). The performance of vector oriented data storage in ESRI's ArcGIS. *Transactions in GIS*, 11:47–65.

Batty, M., Xie, Y. (1991). Model structures, exploratory spatial data analysis, and aggregation. *International Journal of Geographical Information Systems*, 8:291–307.

Bhalla, N. (1991). Object-oriented data models: a perspective and comparative review. *Journal of Information Science*, 17:145–160.

Boguslawski, P., Gold, C.M., Ledoux, H. (2011). Modeling and analysing 3D buildings with a primal/dual data structure. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66:188–197.

Bregt, A.K., Denneboom, J., Gesink, H.J., van Randen, Y. (1991). Determination of rasterizing error: a case study with the soil map of The Netherlands. *International Journal of Geographical Information Systems*, 5:361-367.

Carrara, A., Bitelli, G., Carla, R. (1997). Comparison of techniques for generating digital terrain models from contour lines. *International Journal of Geographical Information Systems*, 11:451–473.

Congalton, R.G. (1997). Exploring and evaluating the consequences of vector-to-raster and raster-to-vector conversion. *Photogrammetric Engineering and Remote Sensing*, 63:425–434.

Downs, R.M. (1998). The geographic eye: seeing through GIS. *Transactions in GIS,* 2:111–121.

Holroyd, F., Bell, S.B.M. (1992). Raster GIS: Models of raster encoding. *Computers and Geosciences*, 18:419–426.

Joao, E.M. (1998). *Causes and Consequences of Map Generalization*. London: Taylor and Francis.

Kumler, M.P. (1994). An intensive comparison of triangulated irregular networks (TINs) and digital elevation models. *Cartographica*, 31:1–99.

Langram, G. (1992). *Time in Geographical Information Systems*. London: Taylor and Francis.

Laurini, R., Thompson, D. (1992). *Fundamentals of Spatial Information Systems*. London: Academic Press.

Lee, J. (1991). Comparison of existing methods for building triangular irregular network models of terrain from grid digital elevation models. *International Journal of Geographical Information Systems*, 5:267–285.

Masser, I. (2005). *GIS Worlds: Creating Spatial Data Infrastructures*. Redlands: ESRI Press.

Nagy, G., Wagle, S.G. (1979). Approximation of polygonal maps by cellular maps. *Communications of the Association of Computational Machinery*, 22:518–525.

Peuker, T.K., Chrisman, N. (1975). Cartographic Data Structures. *The American Cartographer*, 2:55–69.

Peuquet, D.J. (1984). A conceptual framework and comparison of spatial data models. *Cartographica*, 21:66–113.

Peuquet, D.J. (1981). An examination of techniques for reformatting digital cartographic data. Part II: the raster-to-vector process. *Cartographica*, 18:375–394.

Piwowar, J.M., LeDrew, E.F., Dudycha, D.J. (1990). Integration of spatial data in vector and raster formats in geographical information systems. *International Journal of Geographical Information Systems*, 4:429–444.

Rana, S. (2004). *Topological Data Structures for Surfaces: An Introduction to Geographical Information Science*. New York: Wiley.

Rigaux, P., Scholl, M., Voisard, A. (Eds.). (2002). *Spatial Databases: with Application to GIS*. New York: Elsevier.

Shaffer, C.A., Samet, H., Nelson, R.C. (1990). QUILT: a geographic information system based on quadtrees. *International Journal of Geographical Information Systems*, 4:103–132.

Slocum, T.A., McMaster, R.B., Kessler, F.C., Howard, H.H. (2005). *Thematic Cartography and Geographic Visualization*. 2nd ed. New York: Prentice-Hall.

Tomlinson, R.F. (1988). The impact of the transition from analogue to digital cartographic representation. *The American Cartographer,* 15:249–262.

Tuan, A.N.G. (2013). Overview of three-dimensional GIS data models. *International Journal of Future Computer and Communication*, 2:270–274.

Wedhe, M. (1992). Grid cell size in relation to errors in maps and inventories produced by computerized map processes. *Photogrammetric Engineering and Remote Sensing*, 48:1289–1298.

Wilkie, D., Sewall, J., Lin, M.C. (2011). Transforming GIS data into functional road models for large-scale traffic simulation. *IEEE Transactions on Visualization and Computer Graphics*, 18:890–901.

Wise, S. (2002). *GIS Basics*. New York: Taylor & Francis.

Worboys, M.F., Duckham, M. (2004). *GIS: A Computing Perspective*. 2nd ed. Boca Raton: CRC Press.

Zeiler, M. (1999). *Modeling Our World: The ESRI Guide to Geodatabase Design*. Redlands: ESRI Press.

Zhi-Jun, L., Weller, D.E. (2007). A stream network model for integrated watershed modeling. *Environmental Modeling and Assessment*, DOI:10.1007/s10666-007-9083-9.

Zlatanova, S., Rahman, A.A., Pilouk, M. (2002). Trends in 3D GIS Development. *Journal of Geospatial Engineering*, 4:71-80.

Study Questions

2.1 - How is an entity different from a cartographic object?

2.2 - Describe the successive levels of abstraction when representing real-world spatial phenomena on a computer. Why are there multiple levels, instead of just one level in a spatial data representation?

2.3 - Define a data model and describe three primary differences between the two most commonly used data models.

2.4 - Characterize the following lists as nominal, ordinal, or interval/ratio:

a) 1.1, 5.7, -23.2, 0.4, 6.67

b) green, red, blue, yellow, sepia

c) white, light grey, dark grey, black

d) extra small, small, medium, large, extra large

e) forest, woodland, grassland, bare soil

f) 1, 2, 3, 4, 5, 6, 7

2.5 - Characterize the following lists as nominal, ordinal, or interval/ratio:

a) Spurs, Citizens, Reds, Hornets, Baggies, Toffees, Potters

b) pinch, handful, bucket, bushel, truckload

c) 6.2, 7.8, 1.1, 0.5, 19.3

d) gram, kilogram, metric ton

e) Mexico, Canada, Argentina, Guyana, Martinique

f) small, smaller, smallest

2.6 - Indicate which of the following are allowable geographic coordinates:

a) N45° 45’ 45”　　b) longitude -127.34795°　　c) S96° 12’ 33”

d) E 66° 15’ 60”　　e) W -12° 23’ 55”　　f) N 56.9999°

2.7 - Indicate which of the following are allowable geographic coordinates:

a) N145° 45’12”　　b) latitude -62.34795°　　c) S110° 52’ 43”

d) S 49° 15’ 60”　　e) N 89° 59’ 59”　　f) S -46.6000°

2.8 - Convert the following degree measures to radians:

a) 47.2837° b) 155.724° c) -111.2045°

Convert the following radian measures to degrees:

d) 0.0042 e) -1.26 f) 2.25037

2.9 - Convert the following degree measures to radians:

a) 102.83° b) -21.533° c) 92.045°

Convert the following radian measures to degrees:

d) 1.52 e) 0.014 f) 0.37

2.10 - Complete the following coordinate conversion table, converting the listed points from degrees-minutes-seconds (DMS) to decimal degrees (DD), or from DD to DMS. See Figure 2-10 for the conversion formula.

Point	DMS	Decimal Degrees
1	36°45'12"	36.75333
2	114°58'2"	
3	85°19'7"	
4		14.00917
5		275.00001
6		0.99528
7	183°19'22"	

2.11 - Complete the following coordinate conversion table, converting the listed points from degrees-minutes-seconds (DMS) to decimal degrees (DD), or from DD to DMS. See Figure 2-10 for the conversion formula.

Point	DMS	Decimal Degrees
1	97°45'10"	97.75278
2	122°10'2"	
3	15°0'12"	
4		322.19861
5		152.65583
6		5.75
7	23°12'50"	

2.12 - Assume a spherical Earth with a radius of 6378.0 km. Calculate the great circle distances from St. Paul, Minnesota, latitude 44.9537, longitude -93.09 to the following points:

a) Chicago, latitude 41.8781°, longitude -87.6298°

b) Reykjavik, latitude 64.1265°, longitude -21.8174°

c) Buenos Aires, latitude -34.6037°, longitude -58.3816°

2.13 - Assume a spherical Earth with a radius of 6378.0 km. Calculate the great circle distances from St. Paul, Minnesota, latitude 44.9537°, longitude -93.09° to the following points:

a) New York, latitude 40.7128°, longitude -74.0059°

b) Paris, latitude 48.8566°, longitude 2.3522°

c) Tokyo, latitude 35.6895°, longitude 139.6917°

2.14 - What is vector topology, and why is it important? What is planar topology, and when might nonplanar be more useful than planar topology?

2.15 - Identify the number of times each of the following topological rules are broken for the building outlines (red/darker) and property parcels (green/light) polygon layers. Note that all layers are semi-transparent so that you may identify overlaps.

a) Buildings must not overlap.

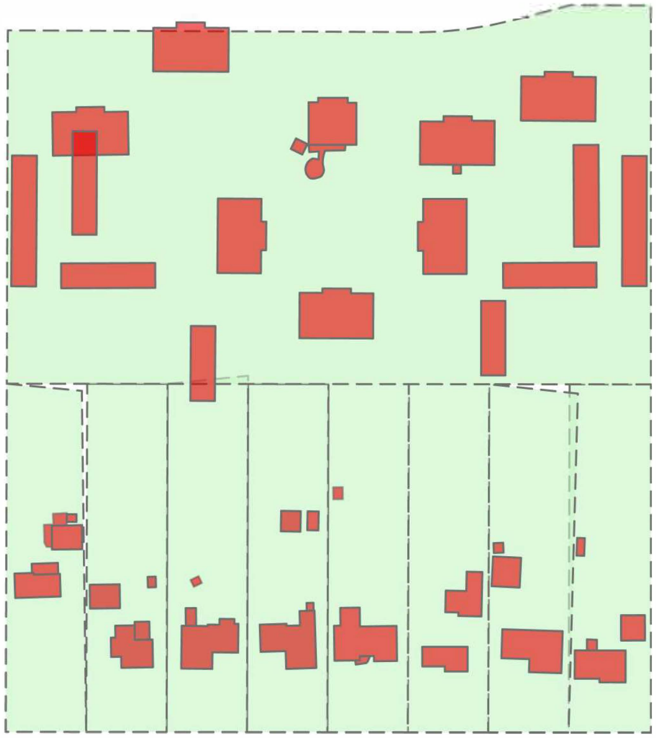
b) Parcels must not overlap.

c) Parcels must not have gaps.

d) Buildings must be entirely within the parcel layer.

e) A building must not span a parcel boundary.

2.16 - Identify the number of times each of the following topological rules are broken for the building outlines (red/darker) and property parcels (green/light) polygon layers. Note that all layers are semi-transparent so that you may identify overlaps.



a) Buildings must not overlap.

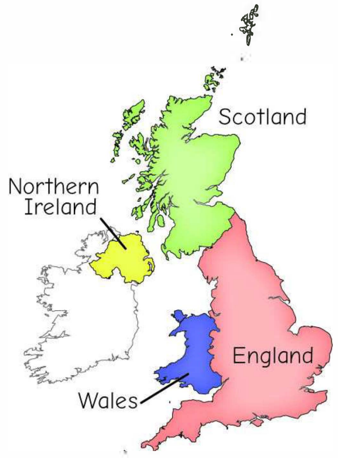
b) Parcels must not overlap.

c) Parcels must not have gaps.

d) Buildings must be entirely within the parcel layer.

e) A building must not span a parcel boundary.

2.17 - Draw multi-part and single-part data layers and tables for the United Kingdom, recording the labeled countries.

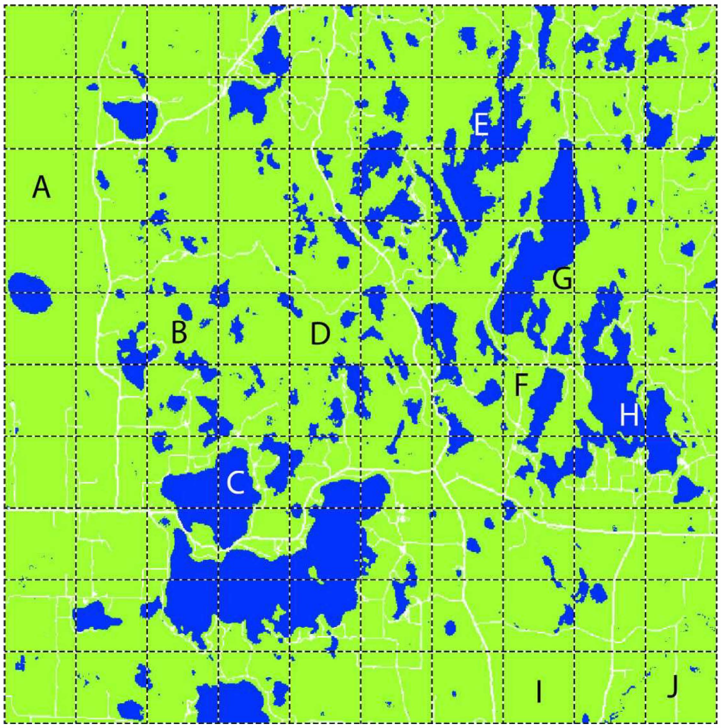


2.18 - Draw multi-part and single-part data layers and tables for the Italian mainland and labeled major islands:.

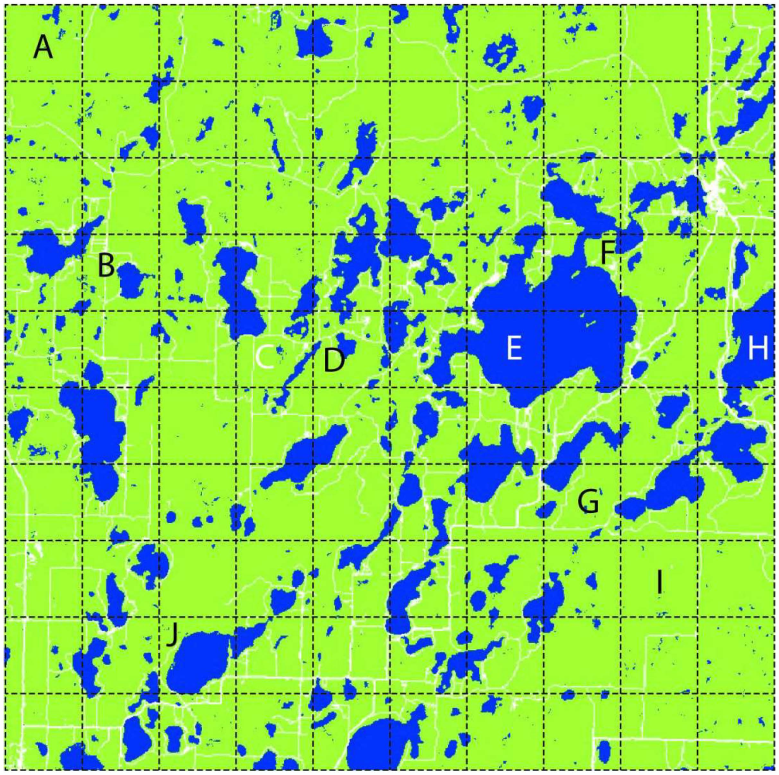
2.19 - What are the respective advantages and disadvantages of vector data models vs. raster data models?

2.20 - Under what conditions are mixed cells a problem in raster data models? In what ways may the problem of mixed cells be addressed?

2.21 - List the labels (A, B, C, etc.) of raster cells that would be assigned as water (dark,blue) and not land (light, green) under a majority coverage rule:



2.22 - List the labels (A, B, C, etc.) of raster cells that would be assigned as water (dark,blue) and not land (light, green) under a majority coverage rule:



2.23 - The following figure shows change in raster resolution, combining four small cells on the left to create an output for each corresponding larger cell on the right. Fill in the two rasters on the right, for the interval/ratio data (top), and the nominal data (bottom). Assume null values are not ignored, and a majority rule for nominal data.

2	2	1	1
2	2	2	2
1	2	1	2
4	5	3	null

a	a	a	a
a	a	a	b
a	b	a	b
c	c	c	d

2.24 - The following figure shows change in raster resolution, combining four small cells on the left to create an output for each corresponding larger cell on the right. Fill in the two rasters on the right, for the interval/ratio data (top), and the nominal data (bottom). Assume null values are not ignored, and a majority rule for nominal data.

2	2	1	2
null	2	2	2
1	null	4	3
3	null	3	4

a	a	b	a
a	b	c	b
null	c	c	c
c	c	c	c

2.25 - Complete the one-to-one (top) and many-to-one (bottom) raster tables in the figure below:

2	2	3	1
2	2	2	2
1	2	1	2
4	5	3	3

→

attribute table
(rows first, start
upper-left corner)

cell-ID	count
2	1
2	1
3	1
⋮	⋮
3	1

2	2	1	1
2	2	2	2
1	2	1	2
4	5	3	3

→

attribute table

cell-ID	count
1	4
2	8
⋮	⋮
5	1

2.26 - Complete the one-to-one (top) and many-to-one (bottom) raster tables in the figure below:

5	7	2	3	3
9	10	4	6	7
8	8	3	4	3
7	7	4	4	3
8	7	4	3	2

→

attribute table
(rows first, start
upper-left corner)

cell-ID	count
5	1
7	1
2	1
⋮	⋮
2	1

5	7	2	3	3
9	10	4	6	7
8	8	3	4	3
7	7	4	4	3
8	7	4	3	2

→

attribute table

cell-ID	count
2	2
3	6
⋮	⋮
10	1

2.27 - What is a triangulated irregular network?

2.28 - What are the main concepts behind object data models, and how do they differ from other data models?

2.29 - Why do we use binary numbers in computers?

2.30 - Express the following base 10 numbers in binary notation:

a) 1	b) 23	c) 256	d) 4
e) 11	f) 10	g) 3	h) 20

2.31 - Express the following base 10 numbers in binary notation:

a) 2	b) 8	c) 9	d) 17
e) 0	f) 128	g) 22	h) 19

2.32 - Express the following binary numbers in base 10 notation:

a) 0101	b) 0001	c) 1111	d) 00101101
e) 1101	f) 1011	g) 10000001	h) 11111111

2.33 - Express the following binary numbers in base 10 notation:

a) 1110	b) 1001	c) 0011	d) 10000101
e) 1000	f) 1010	g) 10010001	h) 11110000

2.34 - Why do we need to compress data? Which are most commonly compressed, raster data or vector data? Why?

2.35 - What are pointers when used in the context of spatial data, and how are they helpful in organizing spatial data?

2.36 - Write the run length coding for each of the rows in this raster:

b	b	a	a	a	c	a	a	a
c	c	b	b	d	d	d	a	a
b	b	b	b	b	b	b	b	b
e	c	f	b	a	d	f	b	a
a	s	a	f	f	f	b	b	a

2.37 - Write the run length coding for each of the rows in this raster:

c	c	c	c	a	a	a	a	a
a	a	b	b	d	d	d	a	a
e	e	e	f	f	f	f	f	e
a	a	a	a	a	a	a	a	a
c	c	a	a	a	b	f	d	e

3 Geodesy, Datums, Map Projections, and Coordinate Systems

Introduction

Geographic information systems are different from other information systems because they include coordinates that define the location, shape, and extent of geographic objects. For effective GIS use, we must clearly understand how coordinate systems are established for the Earth, how coordinates are measured on the Earth's curving surface, and how these coordinates are converted for use in flat maps, either digital or paper. This chapter introduces *geodesy*, the science of measuring the shape of the Earth, and *map projections*, the transformation of coordinate locations from the Earth's curved surface onto flat maps.

Defining coordinates for the Earth's surface is complicated by four main factors. First, most people view geography on a flat surface. We perceive a flat Earth because the curvature is barely perceptible at human scales. We've used flat maps for more than 40 centuries, and although globes are helpful for visualization at extremely small scales, they are impractical for most purposes.

A flat map must distort geometry in some way because the Earth is curved. When we plot latitude and longitude coordinates on a Cartesian system, "straight" lines will appear bent, and polygons will be distorted. This distortion may be difficult to detect on detailed maps that cover a small area, but the distortions become apparent as the mapped area grows. Because measurements on maps are affected by the distortion, we must use a map projection to reconcile the portrayal of the Earth's curved surface onto a flat surface.

The second main problem in defining a coordinate system results from the irregular shape of the Earth. We learn early on that the Earth is shaped as a sphere. This is a valid approximation for many uses, however, it is only an approximation. Past and present natural forces yield an irregularly shaped Earth. This shape affects how we best map the surface of the Earth, and how we define flat coordinate systems.

Third, our measurements are rarely perfect, and this applies when measuring both the shape of the Earth and the exact position of features on it. All locations depend on measurements that contain some error, and on analyses that require assumptions. Our measurements improve through time, and so does the sophistication of our analysis, so our positional estimates improve; this evolution means our estimates of positions change through time.

Finally, the physical locations of points on the Earth change through time. Plate tectonics and vertical crustal movements mean the distance from San Francisco to Tokyo changed from 1950 to 2010, and continues to change today. Earth surface rebound from the weight of past glaciers yields elevations in central Canada several centimeters higher than they were a few

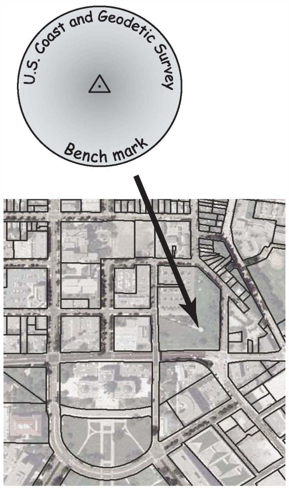
decades ago. How do we specify positions through time when the locations aren't truly fixed?

Because of these four factors, we often have several different sets of coordinates to define the same location on the surface of the Earth. Remember, coordinates are sets of numbers that unambiguously define locations, and in a GIS data layer, we usually use an X (easting), Y (northing) and sometimes height value. But each of these values are only "unique" to any given point for a specified set of measurements, calculation assumptions, at a specified time. The coordinates depend on the reference system we use for measuring latitudes and longitudes (which depends on measurement and Earth's shape), how we translate points from a curved Earth to a flat map surface (which depends on how we project), and to what set of measurements we reference our coordinates (our measurement methods and quality), and when (which depends on crustal movement). We may, and often do, address these factors in a number of different ways, and the coordinates for the same point will be different for these different choices. We can translate between these different choices, as long as we are clear in defining them.

An example may help. Figure 3-1 shows the location of a U.S. survey mark, a precisely surveyed and monumented point. Coordinates for this point are maintained by federal and state government surveyors, and resulting coordinates are shown at the top right of the figure. Note that there are three different versions of the latitude/longitude location for this point. Here, the three versions differ primarily due to differences in the measurements, and how measurement errors were adjusted (the third factor, discussed above). The GIS practitioner may well ask, which latitude/longitude pair should I use? This chapter contains the information that should allow you to choose wisely.

Note that there are also several versions of the x and y coordinates for the point in Figure 3-1. The differences in the coordinate values are too great to be due solely to measurement errors. The differences are due primarily to how we choose to project from the curved Earth to a flat map, and in part to the Earth shape we adopt and the measurement system we use.

Coordinates for a Point Location



From Surveyor Data:

	Latitude (N)	Longitude (W)
NAD83(2007)	44 57 23.23074	093 05 58.28007
NAD83(1986)	44 57 23.22405	093 05 58.27471
NAD83(1996)	44 57 23.23047	093 05 58.27944

	X	Y	
SPC MNS	317,778.887	871,048.844	MT
SPC MNS	1,042,579.57	2,857,766.08	sFT
UTM15	4,978,117.714	492,150.186	MT

From Data Layers:

	X	Y	
MN-Ramsey	573,475.592	160,414.122	sFT
MN-Ramsey	174,195.315	48,893.966	MT
SPC MNC	890,795.838	95,819.779	MT
SPC MNC	2,922,552.206	314,365.207	sFT
LCC	542,153.586	18,266.334	MT

Figure 3-1: An example of different coordinate values for the same point. We may look up the coordinates for a well-surveyed point, and we may also obtain the coordinates for the same point from a number of different data layers. We often find multiple latitude/longitude values (surveyor data, top), or x and y values for the same point (surveyor data, or from data layers, bottom).

Whenever we work with spatial data, we must choose how to address the first three factors: projection distortion, an irregularly shaped Earth, and measurement imprecision. If our data are of very high accuracy and precision and we wish to work across time periods, we must address the fourth factor: vertical and horizontal movements of physical locations through time.

It is crucial to realize that different ways of addressing 1) the Earth's curvature, 2) the Earth's deviation from our idealized shape, 3) inevitable inaccuracies in measurement, and 4) physical shifts, will result in different coordinates. These differences are the root of many errors in spatial analysis. As a rule, you should know the coordinate system used for all of your data, and convert all data to the same coordinate system, for the same time epoch, prior to analysis. In some cases the differences when ignoring some of these four factors may be small in relation to the spatial precision required by your analysis, particularly for the fourth factor (time differences between coordinate measurements). As positioning technology improves, we can make increasingly accurate and precise measurements, so in many cases, the epoch of measurement becomes important. This chapter describes how we define, measure, and convert among coordinate systems.

Modern Coordinate Capture, Coordinate Systems, and Datums

Most GIS data collection relies directly or indirectly on satellite-based positioning systems. These systems, described in detail in Chapter 5, allow the rapid, accurate collection of locations. Positions are referenced to Earth-centered, three dimensional, Cartesian coordinate systems – the X, Y, and Z of 3D systems described in Chapter 2. A specific, defined version of a 3D system is called a *datum*. Datums underpin all geographic measurements. The navigation system operated by the U.S. (GPS) provides coordinates in a datum labeled as WGS84(xxxx), where xxxx represents a version number. In most of North America, collected data are often converted to a different datum, labeled as NAD83(yy) system, where yy is a version number. The other satellite positioning systems (GLONASS, BeiDou, Galileo) typically report in a datum labeled ITRF(zzzz), where zzzz is a version number, usually the year of issue. We will describe WGS84, NAD83, and ITRF datums and how they relate to each other in the first half of this chapter.

These various versions of X, Y, and Z Cartesian coordinates are then commonly converted to latitude, longitude, and height coordinates, and subsequently projected to "flat" coordinate values, suitable for layers in a GIS. This process applies for data directly collected with a GPS or other similar satellite-based navigation system, or with data that depend on satellite positioning, such as satellite or aerial images. Since these coordinate systems differ, have changed through time, and data are commonly converted one to another, it is easy to add error to new data so that features don't fall in their true location. Knowledge of the history and technology of datum development helps us understand how to best collect new data, and to integrate older data with newer measurements.

Early Measurements

In specifying a coordinate system, we must first define the size and shape of the Earth. Humans have long speculated on this. Babylonians believed the Earth was a flat disk floating in an endless ocean, while the Greek Pythagoras, and later Aristotle, reasoned that the Earth must be a sphere. They observed that ships disappeared over the horizon, the moon appeared to be a sphere, and that the stars moved in circular patterns, all consistent with a spherical Earth.

The Greeks next turned toward estimating the size of the sphere. They measured locations on the Earth's surface relative to the Sun or stars, reasoning these provided a stable reference frame. This assumption underlies most geodetic observations taken over the past 2,000 years.

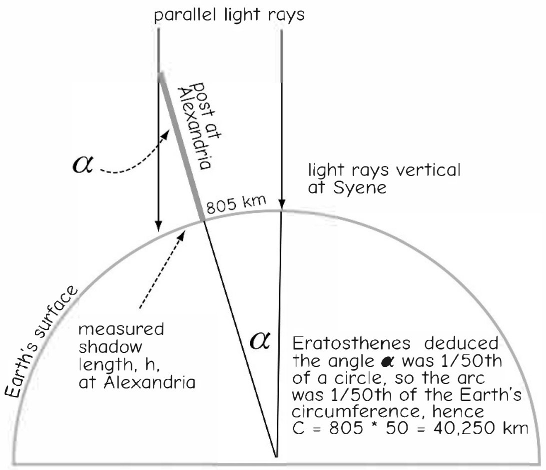


Figure 3-2: Measurements made by Eratosthenes to determine the circumference of the Earth.

Eratosthenes performed early measurements of the Earth's circumference. He noticed that on the summer solstice, the noon sun shone to the bottom of a deep well near the Tropic of Cancer, meaning the sun would be exactly overhead. He also observed that 805 km north, in Alexandria at exactly the same date and time, a vertical post cast a shadow. The shadow/post combination defined an angle that was about $7^{o}12'$, or about 1/50th of a circle (Figure 3-2).

Eratosthenes deduced that the Earth must be 805 multiplied by 50, or about 40,250 kilometers in circumference. His estimate is within 4% of modern measurements of the Earth's circumference.

Specifying the Ellipsoid

By the 18th century, mathematicians argued that centrifugal forces should cause the equatorial regions of the Earth to bulge. They proposed the Earth would be better modeled by an *ellipsoid*, a sphere slightly flattened at the North and South Poles. Expeditions by the French Royal Academy of Sciences starting in 1730 measured the Earth's shape near the Equator and in the high northern latitudes. Complex, repeated, and highly accurate measurements established that an ellipsoid was the best geometric model of the Earth's surface.

Efforts then focused on precisely measuring the size of the Earth's ellipsoid. As noted in Chapter 2, the ellipsoid has two characteristic dimensions (Figure 3-3): the *semi-major axis*, the radius a in the equatorial direction, and the *semi-minor axis*, the radius b in the polar direction. This difference in polar and equatorial radii is also described as a flattening factor, shown in Figure 3-3.

Celestial observations of the stars (Figure 3-4) are combined with long-distance surface measurements to estimate polar and equatorial radii (Figure 3-5). Measurements are repeated over many different locations, and combined for estimates of the semi-major and semi-minor axes. Because early continental surveys could not span most oceans, ellipsoidal parameters were fit for each country, continent, or comparably large survey area.

Measurement efforts through the 19th and 20th centuries led to a set of official ellipsoids which differed in equatorial and polar radii. The Clarke 1866 ellipsoid was commonly used in North America, and was more flattened than the ellipsoid we use today. The Bessel ellipsoid, common in Europe, also specified radii somewhat different than today's best global estimates. Optical instruments predominated before the

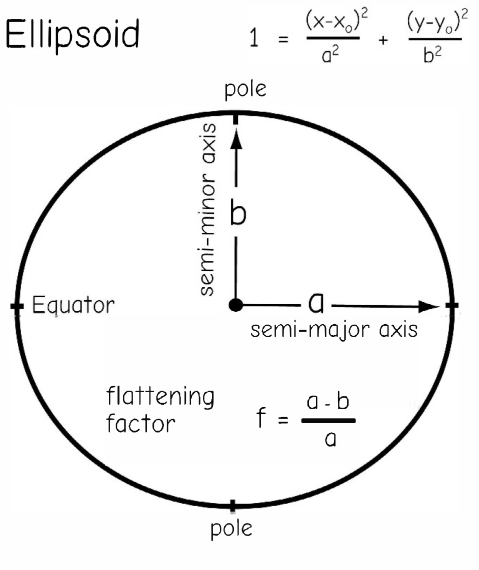


Figure 3-3: An ellipsoidal model of the Earth's shape.

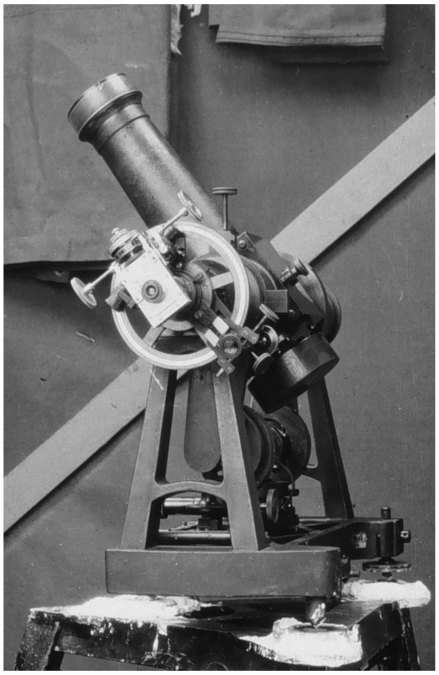


Figure 3-4: An instrument used in the early 1900s for measuring the position of celestial bodies.

early 20th century, and sighting distances were limited by the Earth's curvature. Individual survey legs greater than 50 kilometers (30 miles) were rare, with no good ways to connect surveys across oceans.

Since the 1980s, data derived from satellites, lasers, and broadcast timing signals have been used for extremely precise measurements of relative positions across continents and oceans. Ellipsoids such as the GRS80 provide a "best" overall fit to observed measurements across the globe, and are now preferred and most widely used.

Surface and Ellipsoidal Coordinates

While we make most of our measurements at or near the surface of the Earth, we specify latitudes and longitudes on the ellipsoid, which is usually below the physical surface of the Earth (Figure 3-6). All of our horizontal measurements must be "reduced to," or specified, on the ellipsoid surface. They are mathematically transferred down-

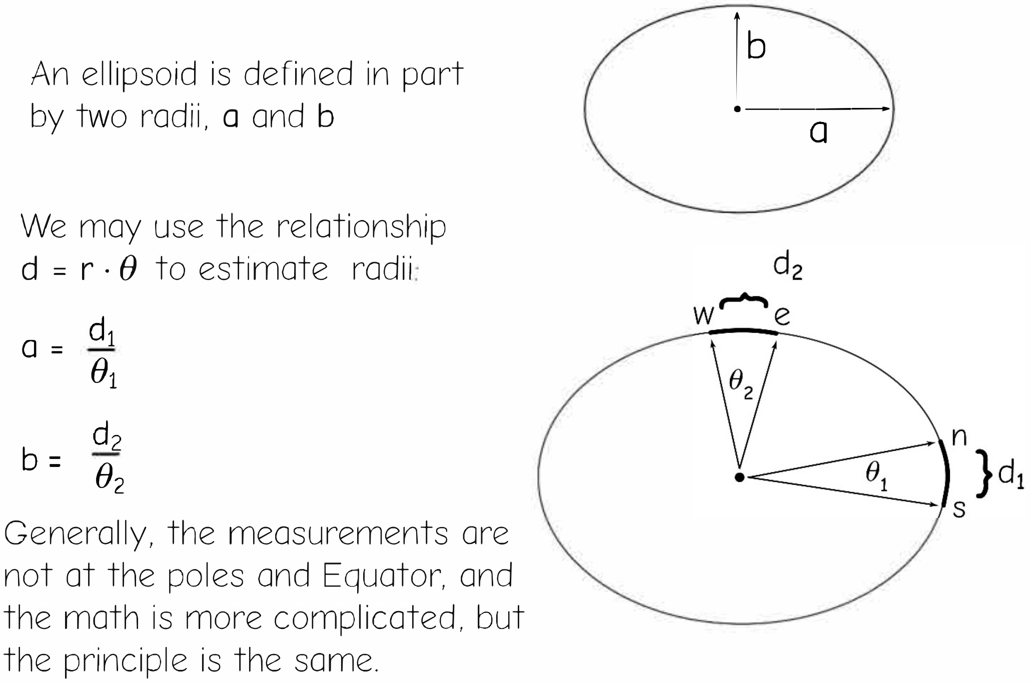


Figure 3-5: Two arcs illustrate the surface measurements and calculations used to estimate the semi-major and semi-minor axes, here for North America. The arc lengths may be measured by surface surveys, and the angles from astronomical observations, as illustrated in Figure 3-2 and Figure 3-3.

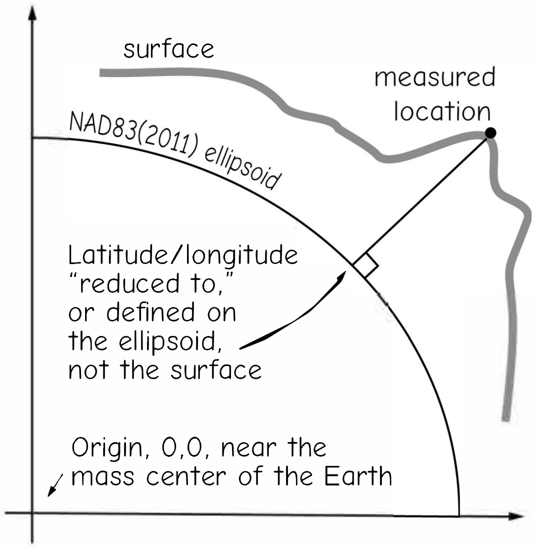


Figure 3-6: Surface measurements are "reduced" downwards onto a chosen ellipsoid directly below the measured location.

ward or upward along a line that is at right angles to the surface of the ellipsoid.

A latitude/longitude location on the ellipsoid is also the latitude/longitude for the surface of the Earth, which may be above or below the ellipsoid. We apply the same latitude and longitude to all points along this line that fall along this right-angle line through the ellipsoid. You can imagine a ray from the ellipsoid up through our surface point. All objects below, on, or above the surface of the Earth and also on this ray will have the same latitude/longitude, for example, a point on the ground and a plane flying directly above that point.

To unambiguously locate an object, e.g., to distinguish a plane in the air from the point on the ground surface directly below it, we must specify a height. We do not use the ellipsoid as a base for our standard heights, and so must introduce another reference surface.

The Geoid

We noted in the introduction that the true shape of the Earth differs slightly from an ellipsoid. Differences in the density of the Earth cause variation in gravitational strength, in turn causing regions to dip below or bulge above a reference ellipsoid (Figure 3-7). This undulating shape is called a *geoid*. In much of the world, including North America, we use a geoid as our zero height.

We define the geoid as the three-dimensional *equipotential surface*, along which the pull of gravity is a specified constant. The geoidal surface may be thought of as an imaginary sea that covers the entire Earth and is not affected by wind, waves, the Moon, or forces other than Earth's gravity. The surface of the geoid extends across the Earth, approximately at mean sea level across the oceans, and continuing under continents at a level set by gravity. The surface is always at right angles to the direction of local gravity.

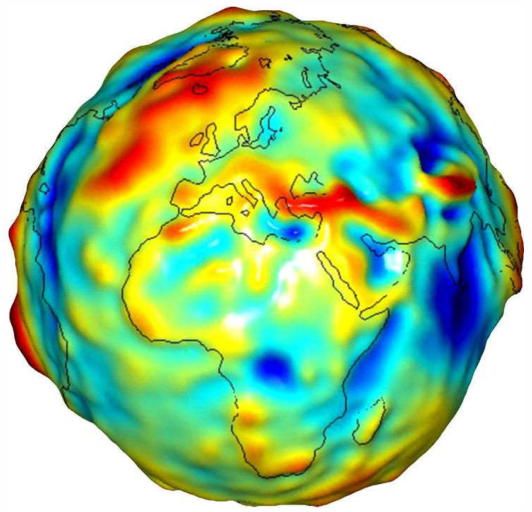


Figure 3-7: Depictions of the Earth's gravity field, as estimated from satellite measurements. These show the undulations, greatly exaggerated, in the Earth's gravity, and hence the geoid (courtesy University of Texas Center for Space Research, and NASA).

We must emphasize that a geoidal surface differs from mean sea level. Mean sea level may be higher or lower than a geoidal surface because ocean currents, temperature, salinity, and wind variations can cause persistent high or low areas in the ocean. These non-gravitational differences can be up to a meter (3 feet), perhaps small on global scale, but large in local or regional analysis. We historically referenced heights to mean sea level, and many believe we still do, but this is no longer true for most spatial data systems.

Because we have two reference surfaces, a geiod and an ellipsoid, we also have two bases from which to measure height. Elevation is typically defined as the distance above a geoid. This elevation above a geoid is also called the *orthometric height* (Figure 3-8), and may be thought of as replacing our older notion of height above mean sea level. Heights above an ellipsoid, or *ellipsoidal heights,* are used in some coordinate system calculations and for some global navigation systems such as GPS, but ellipsoidal heights are not our standard height. These are illustrated in Figure 3-8, with the ellipsoidal height labeled h and orthometric height labeled H. The difference between the ellipsoidal height and orthometric height at any location, shown in Figure 3-8 as N, has various names, including *geoidal height* and *geoidal separation.*

The absolute value of the geoidal height is less than 100 meters over most of the Earth (Figure 3-9). Although it may at first seem difficult to believe, the "average" ocean surface near Iceland is more than 150 meters "higher" than the ocean surface northeast of Jamaica. This height difference is measured relative to the ellipsoid. Since gravity pulls in a direction that is perpendicular to the geoidal surface, the force is at a right angle to the surface of the ocean, resulting in persistent bulges and dips in the mean ocean height. Variation in ocean heights due to swells and wind-driven waves are more apparent at local scales, but are much smaller than the long-distance geoidal undulations.

The geoidal height is quite small relative to the polar and equatorial radii. The Earth's equatorial radius is about 6,780,000 meters, or about 32,000 times the range of the highest to lowest geoidal heights. This small geoidal height is imperceptible at human scales. While relatively small, the geoidal variations in shape must still be considered for accurate vertical and horizontal mapping over continental or global distances.

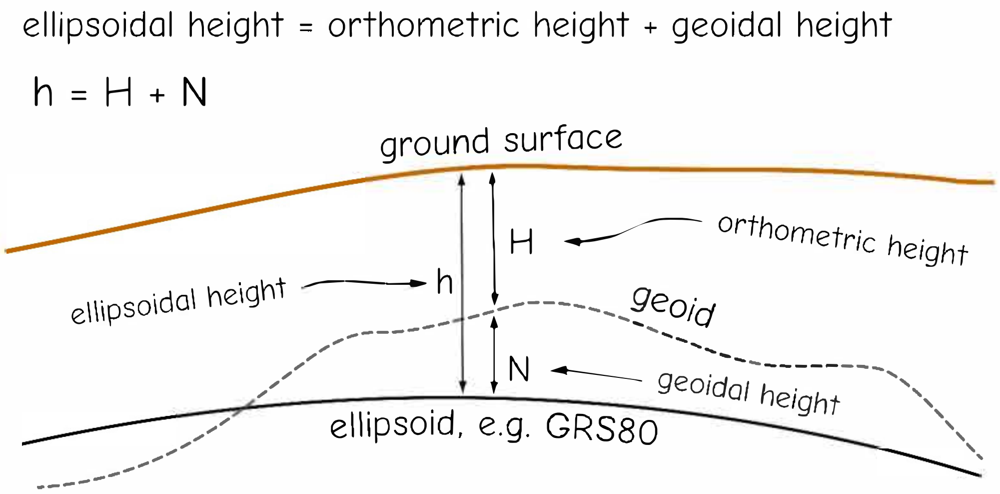


Figure 3-8: Ellipsoidal, orthometric, and geoidal height are interrelated. Note that values for N are highly exaggerated in this figure – values for N are typically much less than H. We often use this formula, e.g., to calculate orthometric height (elevation) when we know the ellipsoidal height (commonly from GPS), and geoidal height (from national models).

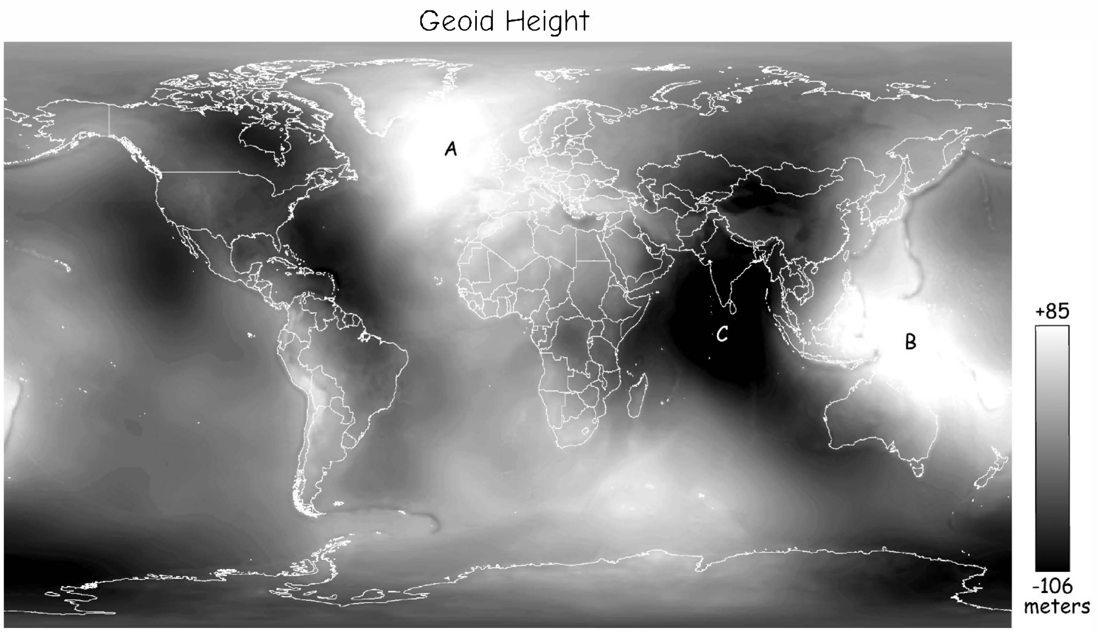


Figure 3-9: Geoidal heights vary across the globe. This figure depicts positive geoidal heights in lighter tones (geoid above the ellipsoid) and negative geoidal heights in darker tones. Note that geoidal heights are positive for large areas near Iceland and the Philippines (A and B, respectively), while large negative values are found south of India (C). Continental and country borders are shown in white.

The geoid is a measured and interpolated surface; unlike an ellipsoid, the geoidal surface is not defined by a simple mathematical equation. The geoid's surface is measured by a number of methods, initially by a combination of *plumb bob*, a weight suspended by a string that indicates the direction of gravity, and horizontal and vertical distance measurements, and later with various types of *gravimeters* (Figure 3-10), devices that measure the gravitational force.

Figure 3-11 shows how differences in the Earth's shape due to geoidal deviations will produce different local ellipsoids. An ellipsoid fit to a local set of points will produce different estimates of the best ellipsoid origin, axis orientation, and ellipsoid radii than surveys that fit points on another part of the Earth. Measurements based on South American surveys yielded a different "best" ellipsoid than those in Europe. Likewise, Europe's best ellipsoidal estimate was different from Asia's, and from South America's, North America's, or those of other regions. One ellipsoid could not be fit to all the world's survey data because during the

Figure 3-10: A portable field gravimeter, an instrument used for measuring gravitational force at a field location. These measurements are combined with surveying measurements to estimate geoidal surfaces (courtesy National Oceanic and Atmospheric Administration, NOAA).

18th and 19th centuries, there was no clear way to combine a global set of measurements.

Satellite-based measurements in the late 20th century substantially improved the global coverage, quality, and density of geoidal height measurements, aiding the development of globally-accurate geoids and ellipsoids. The GRACE experiment, initiated with the launch of twin satellites in 2002, is an example of such improvements. Distances between a pair of satellites are constantly measured as they orbit the Earth. The satellites are pulled closer or drift farther from the Earth due to variation in the gravity field. Because the orbital path changes slightly each day, we eventually have nearly complete Earth coverage of the strength of gravity, and hence the location of the reference gravitational surface. The ESA GOCE satellite, launched in 2009, uses precision accelerometers to measure gravity-induced velocity change. GRACE and GOCE observations have substantially improved our estimates of the gravitational field and geoidal shape.

Satellite and other observations are used by geodesists to develop geoidal models. These support a series of geoid estimates, for example, by the U.S. NGS with GEOID90 in 1990, with succeeding geoid estimates in 1993, 1996, 1999, 2003, 2009, and 2012. These are called models because we measure geoidal heights at points or along lines at various parts of the globe, but we need geoidal heights everywhere. Equations are statistically fit that relate the measured geoidal heights to geographic coordinates. Given any set of geographic coordinates, we may then estimate the geoidal height. These models provide an accurate estimation of the geoidal heights for the entire globe.

Horizontal Datums

The geographic coordinate system described in Chapter 2 is based on an established zero meridian passing near the Greenwich Observatory, in England. However, this

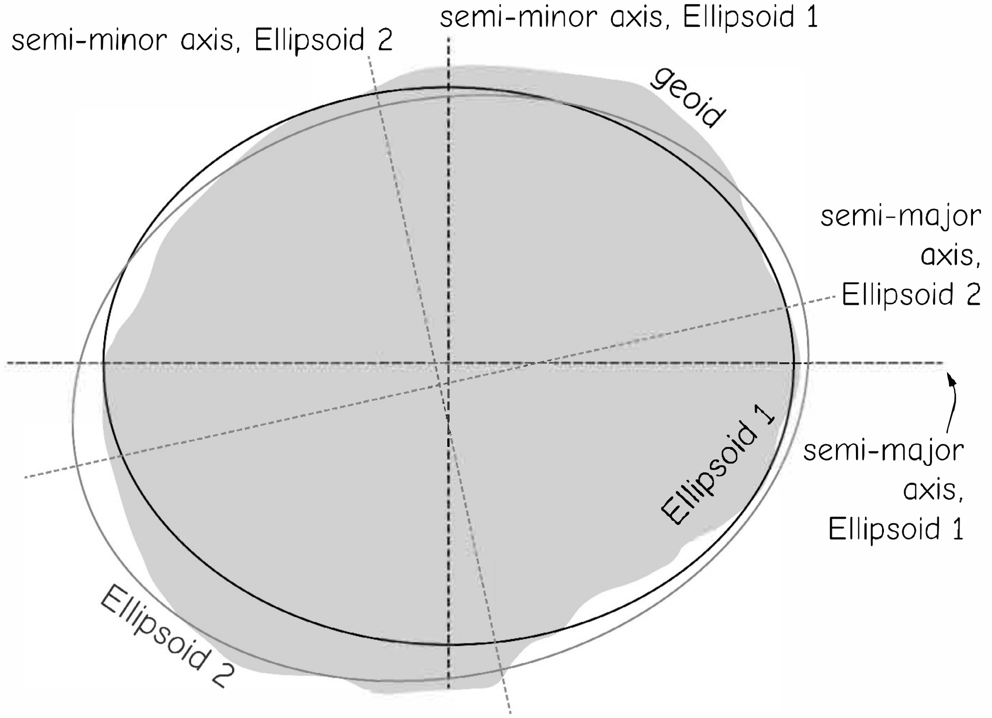


Figure 3-11: Different ellipsoids were estimated due to local irregularities in the Earth's shape. Local best-fit ellipsoids varied from the global best fit, but until the 1970s, there were few good ways to combine global geodetic measurements.

gives us the exact longitude of only one arc, the zero line of longitude. We must estimate the longitudes and latitudes of all other locations through surveying measurements, until recently by observing stars and by measuring distances and directions between points. These surveying methods have since been replaced by modern, satellite-based positioning, but even these new methods are ultimately dependent on astronomical observations. Through these methods, we establish a set of points on Earth for which the horizontal and vertical positions have been accurately determined. These accurately determined points and associated measured and mathematical surfaces are *datums*, references against which we measure all other locations.

These well-surveyed points allow us to specify a *reference frame*, including an origin or starting point. If we are using an ellipsoidal reference frame, we must also specify the orientation and radii of our ellipsoid. If we are using a three-dimensional Cartesian reference frame, we must specify the X, Y, and Z axes, including their origin and orientation. We can choose different values for these various parts of our reference frame, and hence can have different reference frames. All other coordinate locations we use are measured with reference to the chosen reference frame. We then must painstakingly measure a precise set of highly accurate points, so we can express locations relative to this reference frame. For most of the past 150 years, the most accurate observations were referenced to the Sun, stars, or other celestial bodies (Figure 3-12), as they provided the most stable way to establish our reference frame.

Figure 3-12: Astronomical observations were used in early geodetic surveys to measure datum locations (courtesy NOAA).

Many countries have a government body charged with making precise surveys of points to help define this reference frame, and make the frame useful to users. For example, most surveys in the United States are related back to high accuracy points maintained by the National Geodetic Survey (NGS). The NGS establishes geodetic latitudes and longitudes of known points, most of which are monumented with a metal disk, concrete posts, or other durable markers.

A *geodetic datum* is a reference surface. A geodetic datum consists of two major components. The first component is an ellipsoid with a spherical or three-dimensional Cartesian coordinate system and an origin. Eight parameters are needed to specify the

Figure 3-13: A bronze disk used to monument a survey mark.

ellipsoid: a and b to define the size/shape of the ellipsoid; the X, Y, and Z values of the origin; and an orientation angle for each of the three axes.

A datum includes a set of positions that have been painstakingly surveyed, against which subsequent surveys are referenced. A datum is sometimes defined as a reference surface, and a *realization of a datum* as that surface plus a network of precisely measured points. The measured points describe a *Terrestrial Reference Frame,* or specific measured datum. This clearly separates the theoretical reference surface from a useful terrestrial reference frame, complete with points from which we can survey new points or re-measure old. While this more precise language may avoid some confusion, datum commonly refers to both the defined surface and the various realizations of each datum.

Precisely surveyed points are commonly known as *survey marks* and *bench marks*, with the latter often reserved for precise vertical surveys. Marks often consist of a metal disk embedded in rock or concrete (Figure 3-13), although they also may consist of marks chiseled in rocks, embedded iron posts, or other long-term marks. Due to the considerable effort and cost of establishing the coordinates for each survey mark, they are often redundantly monumented, and their distance and direction from specific local features are recorded. Control survey points are often identified with a number of nearby signs to aid in recovery (Figure 3-14).

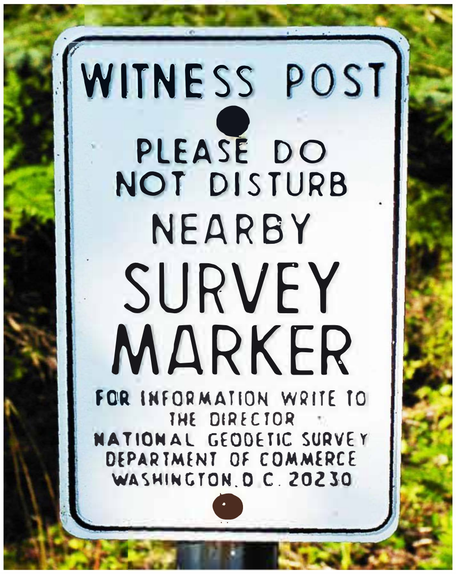


Figure 3-14: Signs are often placed near control points to warn of their presence and aid in their location.

The NGS maintains and disseminates information on survey marks in the United States (Figure 3-15), with access via the World Wide Web (http://www.ngs.noaa.gov). Stations may be found based on a station name, a state and county name, a type of station (horizontal or vertical), by survey order, survey accuracy, date, or coordinate location. These stations may be used as reference points to check the

```
National Geodetic Survey,   Retrieval Date = SEPTEMBER 26, 2011
OB0554  DESIGNATION -  CAPE SMALL  OB0554  PID        -  OB0554
OB0554  STATE/COUNTY-  ME/SAGADAHOC   USGS QUAD   -  PHIPPSBURG (1957)
OB0554
OB0554                        *CURRENT SURVEY CONTROL
OB0554  ____________________________________________________________________
OB0554* NAD 83(1996)    -  43 46 42.87649(N) 069 50 42.26065(W)    ADJUSTED
OB0554* NAVD 88         -        73.     (meters)  240.    (feet)  SCALED
OB0554  ____________________________________________________________________
OB0554  LAPLACE CORR-          2.33  (seconds)                     DEFLEC99
OB0554  GEOID HEIGHT-        -25.73  (meters)                      GEOID03
OB0554  HORZ ORDER  -  FIRST
```

Figure 3-15: A portion of a National Geodetic Survey control point data sheet.

accuracy of any data collection method, for example, new GPS/GNSS equipment, or as a starting point for additional surveys.

Different datums are specified through time because our realizations, or estimates of the datum, change through time. New points are added and survey methods improve. We periodically update our datum when there are enough new or better measurements of survey points, or when we change the parameters of the reference frame (e.g., origin, ellipsoid shape). We do this by reestimating the coordinates of our datum points after including these changes, thereby improving our estimate of the position of each point.

There are two main eras of datums, those created before satellites geodesy, and those after. Satellite positioning technologies became commonplace in the last decade of the 20th century, and substantially increased the number and accuracy of datum points. Datums and coordinates found today are a mix of those developed under pre-satellite datums, and those referenced to post-satellite datums, so the GIS user should be familiar with both.

Geodetic surveys in the 18th and 19th centuries combined horizontal measurements with repeated, excruciatingly precise astronomical observations. Astronomical observations were typically used at the starting point, a few intermediate points, and near the end of geodetic surveys. Astronomical positioning required repeated measurements over several nights. Clouds, haze, or a full moon often lengthened the measurement times. In addition, celestial measurements required correction for atmospheric refraction, a process that bends light and changes the apparent position of stars.

Historically, horizontal optical surveys were as precise and much faster than astronomical methods when measuring over distances up to several tens of kilometers. These horizontal surface measurements were used to connect astronomically surveyed points and thereby create an expanded, well-distributed set of known datum points. Figure 3-16 shows an example survey, where open circles signify points established by astronomical measurements and filled circles denote points established by surface measurements.

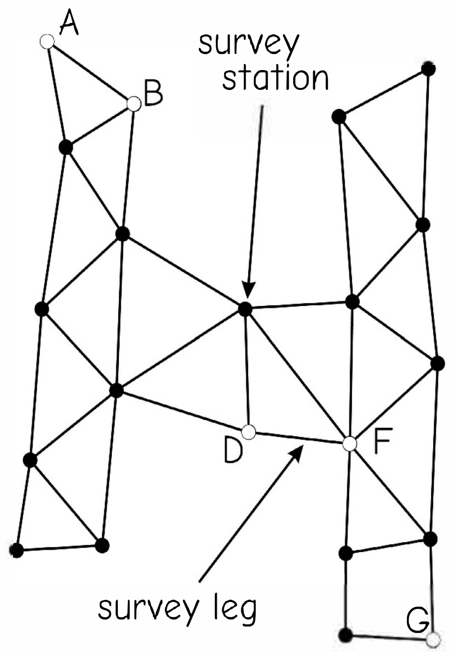


Figure 3-16: A triangulation survey network. Stations may be measured using astronomical (open circles) or surface surveys (filled circles).

Figure 3-16 also illustrates a *triangulation survey*, commonly used prior to satellite positioning. They employ a network of interlocking triangles to determine positions at survey stations. Triangulation surveys were adopted because we can create them through optical angle measurement, with few surface distance measurements, an advantage in the late 18th and early 19th centuries when many datums were first developed. Triangulation also improves accuracy; because there are multiple measurements to each survey station, the location at each station may be computed by various paths.

Triangulation networks spanned long distances, from countries to continents (Figure 3-17). Individual measurements of these triangulation surveys were rarely longer than a few to tens of kilometers; however, each leg of the larger triangles were made up themselves of smaller triangulation traverses.

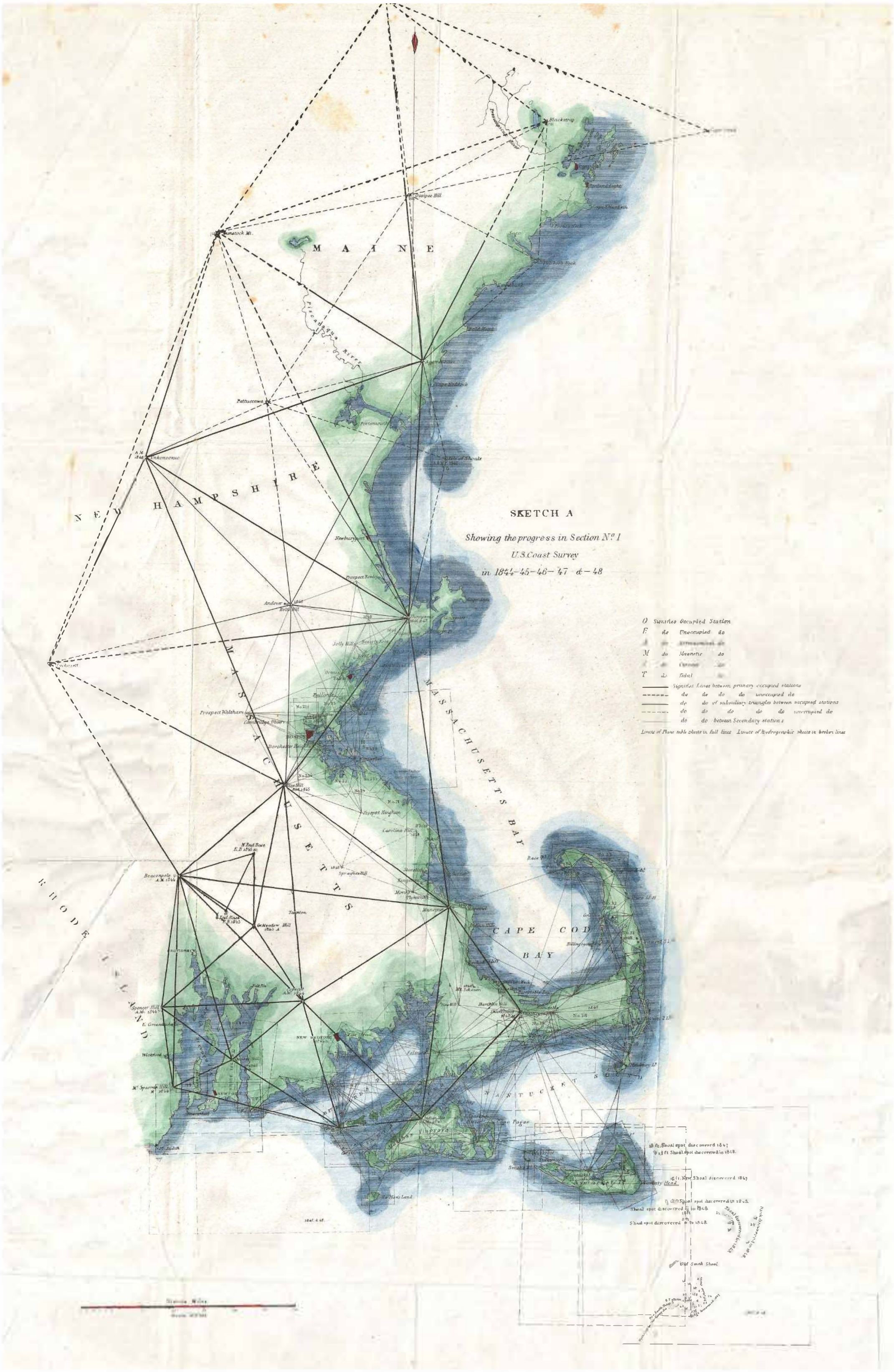


Figure 3-17: A map of the triangulation survey network established the northeastern coast of the United States of America in the early 1800s. Each leg of the triangles, shown here as a single line, is in turn a triangulation survey. This nested triangulation provides reinforcing measurements, thereby increasing the accuracy of the surveyed positions (courtesy NOAA).

Datum Adjustment

Once a sufficiently large set of points has been surveyed, the survey measurements must be harmonized into a consistent set of coordinates. Small inconsistencies are inevitable in any large set of measurements, causing ambiguity in locations. In addition, the long reaches spanned by the triangulation networks, as shown in Figure 3-17, could be helpful in recalculating certain constants, such as the Earth's curvature (see Figure 3-5), which in turn affect the calculations of each surveyed location. The positions of all points in a reference datum are estimated in a network-wide *datum adjustment*. The datum adjustment reconciles errors across the network, first by weeding out blunders or obvious mistakes, and also by mathematically minimizing errors by combining repeat measurements and statistically assigning higher influence to more precise measurements. A datum adjustment only incorporates measurements up to a given point in time, and may be viewed as our best estimate, at that point, of the measured set of locations.

Periodic datum adjustments result in series of regional or global reference datums. Each datum is succeeded by an improved, more accurate datum. This is not a trivial exercise, considering the adjustment may include survey data for tens of thousands of old and newly surveyed points from across the continent, or even the globe. Because of their complexity, these continent-wide or global datum calculations were once infrequent. Computers have improved such that datum adjustments now occur every few years.

A datum adjustment usually results in a change in the coordinates for all existing datum points, as coordinate locations are estimated for both old and new datum points. Our best estimates of the datum point coordinates will change. Differences between the datums reflect differences in the

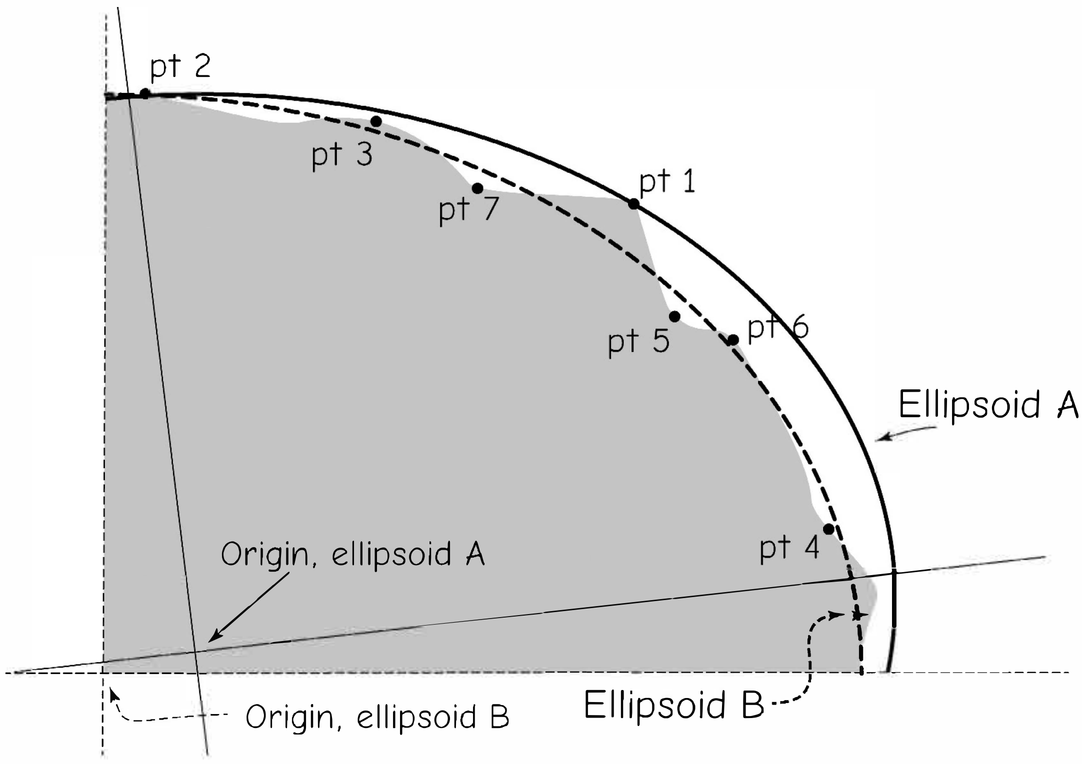


Figure 3-18: An illustration of two datums, one corresponding to Ellipsoid A and based on the fit to pt1 and pt2, and a subsequent datum resulting in Ellipsoid B, and based on a fit of pt1 through pt7. As the number and quality of our survey data improve, subsequent estimates of our best-fitting ellipsoid change.

control points, survey methods, mathematical models, and assumptions used in the datum adjustment.

Figure 3-18 illustrates how ellipsoids might change over time, mostly in the origin and orientation in this example, even for the same survey region. Ellipsoid A is estimated with the datum coordinates for pt1 and pt2, with the shown corresponding coordinate axes, origin, and orientation. Ellipsoid B is subsequently fit, after pts 3 through 7 have been collected. This newer ellipsoid has a different origin and orientation for its axis, causing the coordinates for pt1 and pt2 to change. The points have not moved, but the best estimate of their locations will have changed, relative to the origin set by the new, more complete set of datum points. You can visualize how the latitude angle from the origin to pt1 will change because the origin for ellipsoid A is in a different location than the origin for ellipsoid B. This apparent, but not real, movement is called the *datum shift*, and is expected with datum adjustments.

Commonly Used Datums

Three main series of horizontal datums have been used widely in North America. The first of these is the NAD series, beginning with the *North American Datum of 1927* (NAD27). NAD27 is a legacy datum, still encountered with some older data. NAD27 was a general least squares adjustment that used the Clarke Ellipsoid of 1866 and held fixed the latitude and longitude of a survey station in Kansas.

The *North American Datum of 1983* (NAD83) is the successor to NAD27. We place a modifier in parentheses after the NAD83 designator, e.g., NAD83(1986) to indicate the year, or version, of the datum adjustment. The original NAD83(1986) included approximately 250,000 stations and 2,000,000 distance measurements. The GRS80 ellipsoid was used, an Earth-centered reference, rather than fixing a surface station as with NAD27. Coordinate shifts from NAD27 to NAD83(1986) were large, often tens to 100 meters. In most instances, the surveyed points physically moved very little, for example, due to tectonic plate shifts, but our best estimates of point location changed.

Precise satellite positioning data became widely available soon after the initial NAD83(1986) adjustment, and were often more accurate than NAD83(1986) position estimates. Between 1989 and 2004, the NGS collaborated with other organizations to create *High Accuracy Reference Networks* (HARNs), also known as *High Precision Geodetic Networks* (HPGN) for most of the U.S. Generally, there is a different NAD83(HARN) for each state or small groups of states.

The HARN and subsequent NAD83 adjustments are largely satellite-based, and mark the transition from physical and optical surveying to GPS/GNSS surveying. They underpin a network of Continuously Operating Reference Stations (CORS, Figure 3-19). The CORS network of satellite observations allowed improved datum realizations, e.g. NAD83(CORS93), NAD83(CORS94), NAD83(CORS96), NAD83(2007), and NAD83(2011). NAD83(2011) is a long-

Figure 3-19: A Continuously Operating Reference Station (CORS), used to collect high-accuracy positional measurements from satellites for modern datum development (courtesy NOAA).

observation adjustment based on CORS stations, with coordinates reestimated for a broad set of survey marks. Both the CORS stations themselves and the bench marks are often used as starting points for more precise local surveys.

The *World Geodetic System of 1984* (WGS84) is a second set of datums developed and primarily used by the U.S. Department of Defense (DOD). It was introduced in 1987 based on Doppler satellite measurements of the Earth, and is used in most DOD maps and positional data. The WGS84 ellipsoid is similar to the GRS80 ellipsoid. WGS84 has been updated with more recent satellite measurements and is specified using a version designator. The update based on data collected up to January 1994 is designated as WGS84(G730). WGS84 datums are not widely used outside of the military because they are not tied to a set of broadly accessible, documented physical points.

There have been several subsequent WGS84 datum realizations. The original datum realization exhibited positional accuracy of key datum parameters to within one to two meters. Subsequent satellite observations improved accuracies. A reanalysis was conducted on data collected through week 873 of the GPS satellite schedule, resulting in the more accurate WGS84(G873). Successive realizations are known as WGS84(G1150), WGS84(1674), and WGS84(G1762), and there will likely be more adjustments in the future.

The third set of datums, commonly used worldwide and increasingly in North America, is known as the *International Terrestrial Reference Frames* (ITRF), with datum realizations of the International Terrestrial Reference System (ITRS). A primary purpose for ITRS is to estimate continental drift and crustal deformation by measuring the location and velocity of points, using a worldwide network of measurement locations. Each realization is noted by the year, for example, ITRF89, ITRF90, ITRF91. Each includes the X, Y, and Z location of each point and the velocity of each point in three dimensions. The European Terrestrial Reference System datum (ETRS89 and frequent updates thereafter) is based on ITRF measurements.

The ITRF and WGS84 datums are maintained by different organizations and based on different sets of measurements, but they have been aligned since 1995, and can be considered equivalent for most purposes, as differences between them since 1995 are generally only a few centimeters.

Although they are both based on modern satellite and other accurate measurements, the ITRF and current NAD83 datums do not align, and coordinates can be off by as much as two meters. Since the WGS84 is aligned with the ITRF series, WGS84 also differ by as much as two meters from NAD83. We should be careful in correctly adjusting for datum shifts between the ITRF/WGS84 and NAD83 datums.

Figure 3-20 illustrates the relative size of datum shifts at an NGS marks between various versions of the NAD, and a WGS84/ITRF, based on estimates provided by the National Geodetic Survey. Notice that the datum shift between NAD27 and NAD83(86) is quite large, approximately 40 meters (130 feet), typical of the up to hundreds of meters of shifts from early, regional datums to modern, global datums. The figure also shows the subsequently smaller shifts for NAD83 datums through time, and relatively larger distance between NAD83 and WGS84/ITRF datums.

A datum shift does not imply that points have moved. Most monumented points are stationary relative to their immediate surroundings. The locations change over time as the large continental plates move, but these changes are small, on the order of a few millimeters per year, except in tectonically active areas such as coastal California; for most locations, it is just our estimates of the coordinates that have changed. As survey measurements improve through time and there are more of them, we obtain better estimates of the true locations of the monumented datum points.

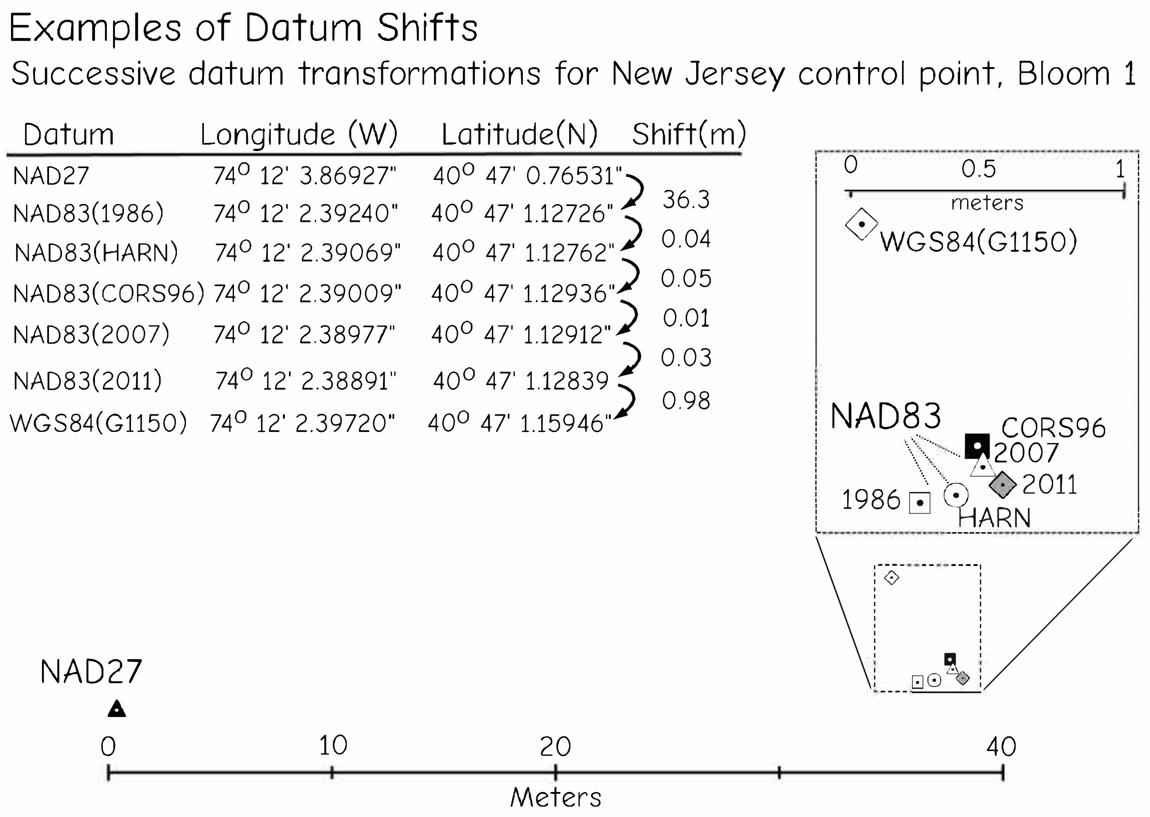


Figure 3-20: Datum shifts in the coordinates of a point for some common datums. Note that the estimate of coordinate position shifts approximately 36 m from the NAD27 to the NAD83(1986) datum, while the shift from NAD83(1986) to NAD83(HARN) then to NAD83(CORS96) are 0.05 m or less. The shift to WGS84(G1150) is also shown, here approximately 0.98 m. Note that the point may not be moving, only our datum estimate of the point's coordinates. Calculations are based on NGS data sheets, NCAT, and HTDP software.

We must emphasize while much data are collected in WGS84/ITRF datums using GNSS (such as GPS), most data are converted to a local or national datum before use in a GIS. In the United States, this typically involves GNSS accuracy augmentation, often through a process called differential correction, described in detail in Chapter 5. Corrections are often based on an NAD83 datum, effectively converting the coordinates to the NAD83 reference, but ITRF datums are also commonly used. Ignorance of this "implicit" conversion among datums is a common source of error in spatial data, and should be avoided.

There are a few points about datums that must be emphasized. First, different datums specify different coordinate systems. You do not expect coordinates for any physical point to be the same when they are expressed relative to different datums.

Second, the version of the datum is important. NAD83(1996) is a different realization than NAD83(2011), and ITRF88 is different than ITRF05. The datum is incompletely specified unless the version is noted. Many GIS software packages refer to a datum without the version, for example, NAD83. This is indeterminate, and confusing, and shouldn't be practiced. It forces the user to work with ambiguity.

Third, differences between families of datums change through time. The NAD83(1986) datum realization is up to two meters different than the NAD83(CORS96), and the original WGS84 differs from the current WGS84 version by more than a meter over much of the Earth. Differences in datum realizations depend on the versions and location on Earth. This means you should assume all data should be converted to the same datum and version before combi-

nation in a GIS. This rule may be relaxed only after you have verified that the datum difference errors are small compared to other sources of error, or small compared to the data accuracy required for the intended spatial analysis.

The U.S. is developing the successor to the NAD83 system, to replace it with the *North American Terrestrial Reference Frame of 2022* (NATRF2022). Most of the differences between this new datum and the ITRF/WGS84 datums will be resolved. The ITRF uses the most current estimates for the mass center of the Earth as the ellipsoid origin, while the NAD83 maintained an earlier, less accurate mass center across the 1986 through present versions (Figure 3-21). This choice was made to postpone the confusion inherent with calculating new coordinates for the huge number of survey marks across the country. In the United States, most spatial data are tied to the widely distributed set of surveyed and marked points reported in the NAD83(CORSxx) datums, and state, county, and local surveys are referenced to these points. The adoption of NATRF2022 will require transforming current NAD83(2011) and earlier data to new coordinates, but will help avoid substantial confusion in position due to datums.

Datum Transformations

Converting coordinates from one datum to another typically requires a *datum transformation*. A datum transformation provides the latitude and longitude of a point in one datum when we know them in another datum; for example, we can calculate the latitude and longitude of a survey mark in NAD83(2011) when we know these geographic coordinates in ITRF08 (Figure 3-22).

Datum transformations are often more complicated when they involve older datums. Many older datums were created piecemeal to optimize fit for a country or continent, so simple formulas often do not exist for transformations involving many older datums, for example, from NAD27 to NAD83. Specialized datum transformations may be provided, usually by government agencies. As an example, in the United States, the National Geodetic Survey created NCAT, a datum transformation tool to convert between various NAD datums.

Transformation among newer datums may use more general mathematical transformations between three-dimensional, Cartesian coordinate systems (Figure 3-22). Transformation equations allow conversion among most NAD83, WGS84, and ITRF

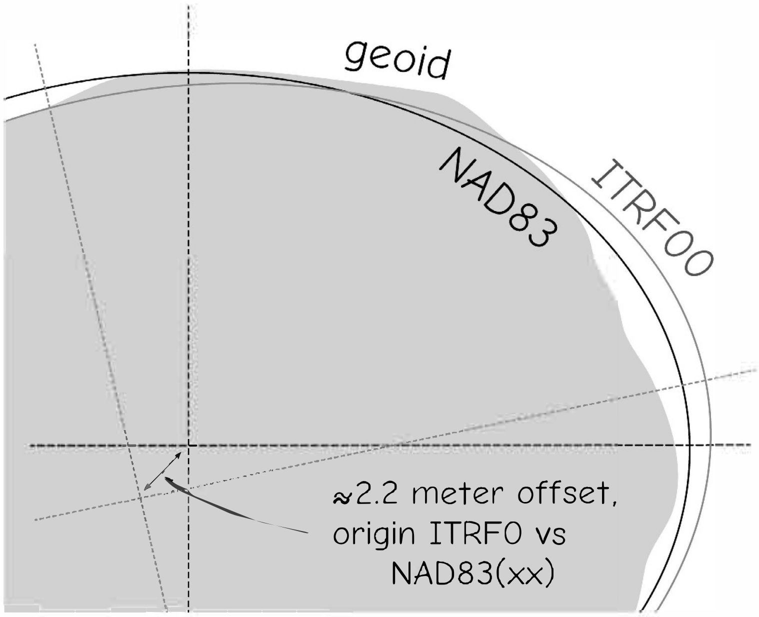


Figure 3-21: The NAD83 and ITRF datums use similar ellipsoid diameters, but different ellipsoid origins and orientations, so coordinates will change when transformed between them. The xx in NAD83(xx) indicates this offset is present through all versions of the NAD83 datum.

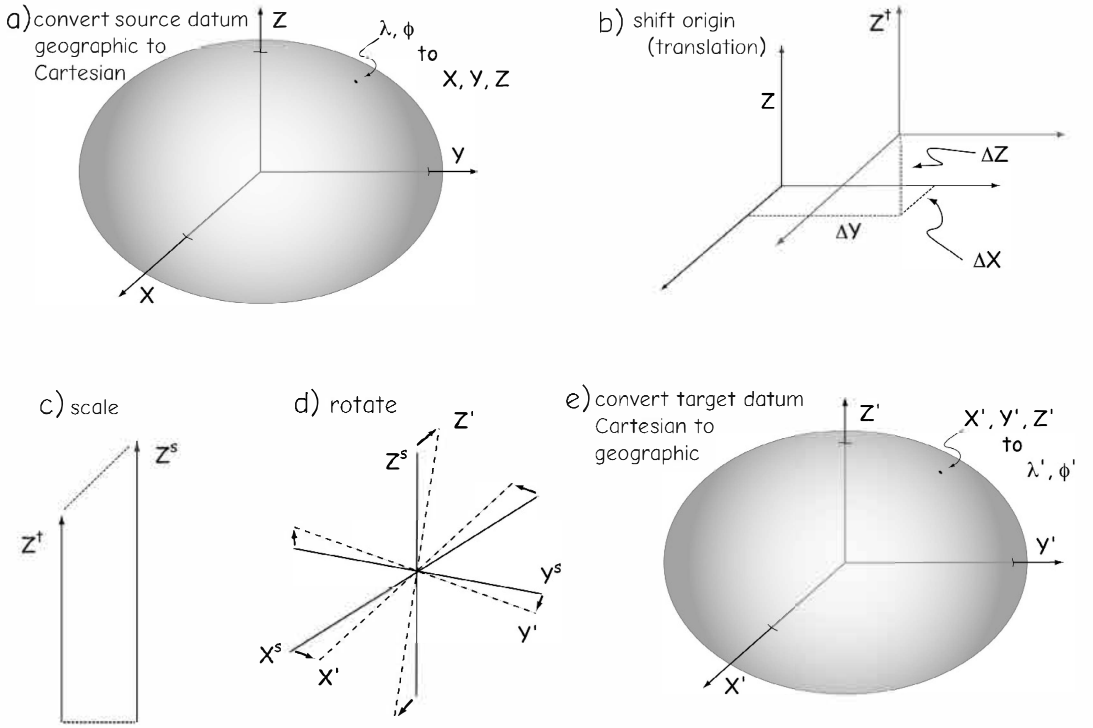


Figure 3-22: Application of a modern datum transformation. Geographic coordinates (longitude, λ, and latitude, ϕ), are transformed to a new datum by a) conversion from geographic to Cartesian coordinates in the old datum (through a set of equations that are not shown), b) applying an origin shift, c) scaling, d) rotating these shifted coordinates, and e) converting these target datum Cartesian coordinates, X', Y', Z', to the longitude and latitude, λ', ϕ', in the target datum.

systems, and are supported in large part by improved global measurements from satellites, as described in the previous few pages. This approach incorporates a shift in the origin, a rotation, and a change in scale from one datum to another.

A datum transformation is typically a multi-step process. In past times, empirical, grid-based methods have been used because many early datums were not strictly derived from coherent mathematical surfaces. Later, a *Molodenski transformation* was common, using a system of equations with three or five parameters. More currently, a *Helmert transformation* is employed using seven or 14 parameters (Figure 3-22). First, geographic coordinates on the source datum are converged from longitude (λ) and latitude (ϕ) to X, Y, and Z Cartesian coordinates. An origin shift (translation), rotation, and scale are applied. This system produces new X', Y', and Z' coordinates in the target datum. These X', Y', and Z' Cartesian coordinates are then converted back to the longitudes and latitudes (λ' and ϕ'), in the target datum.

More advanced methods allow these seven transformation parameters to change through time, to account for tectonic and other shifts, for a total of 14 parameters. These methods are incorporated into software that calculate transformations among modern datums, for example, the Horizontal Time Dependent Positioning (HTDP) tool available from the U.S. NGS (www.ngs.noaa.gov/TOOLS/Htdp/Htdp.shtml). HTDP converts among recent NAD83 datums and most ITRF and WGS84 datums.

Because of tectonic plate movement, the most precise geodetic measurements refer to the epoch, or fixed time period, at which the point was measured or datum fit. The HTDP software includes options to calculate the shift in a location due to different reference datums [for example, NAD83(CORS96) to WGS84(G1150)], the shift due to different

realizations of a datum [for example, NAD83(CORS96) to NAD83(2011)], the shift due to measurements in different epochs [for example, NAD83(CORS96) epoch 1997.0 to NAD83(CORS96) epoch 2010.0], and the differences due to all three factors. Since most points are moving at velocities less than 0.1 mm per year in the NAD83 reference frame, epoch differences are often ignored for all but geodetic surveys.

Datums shifts associated with datum transformations have changed with each successive datum realization, as summarized in Figure 3-23. Several datum pairs are considered equivalent for many purposes when combining data from different data layers, or when applying datum transformations. The WGS84(G730) was aligned with the ITRF92 datum, so these may be substituted in datum transformations requiring no better than centimeter accuracies. Similarly, the WGS84(G1150) and ITRF00 datums have been aligned, and may be substituted in most subsequent transformations.

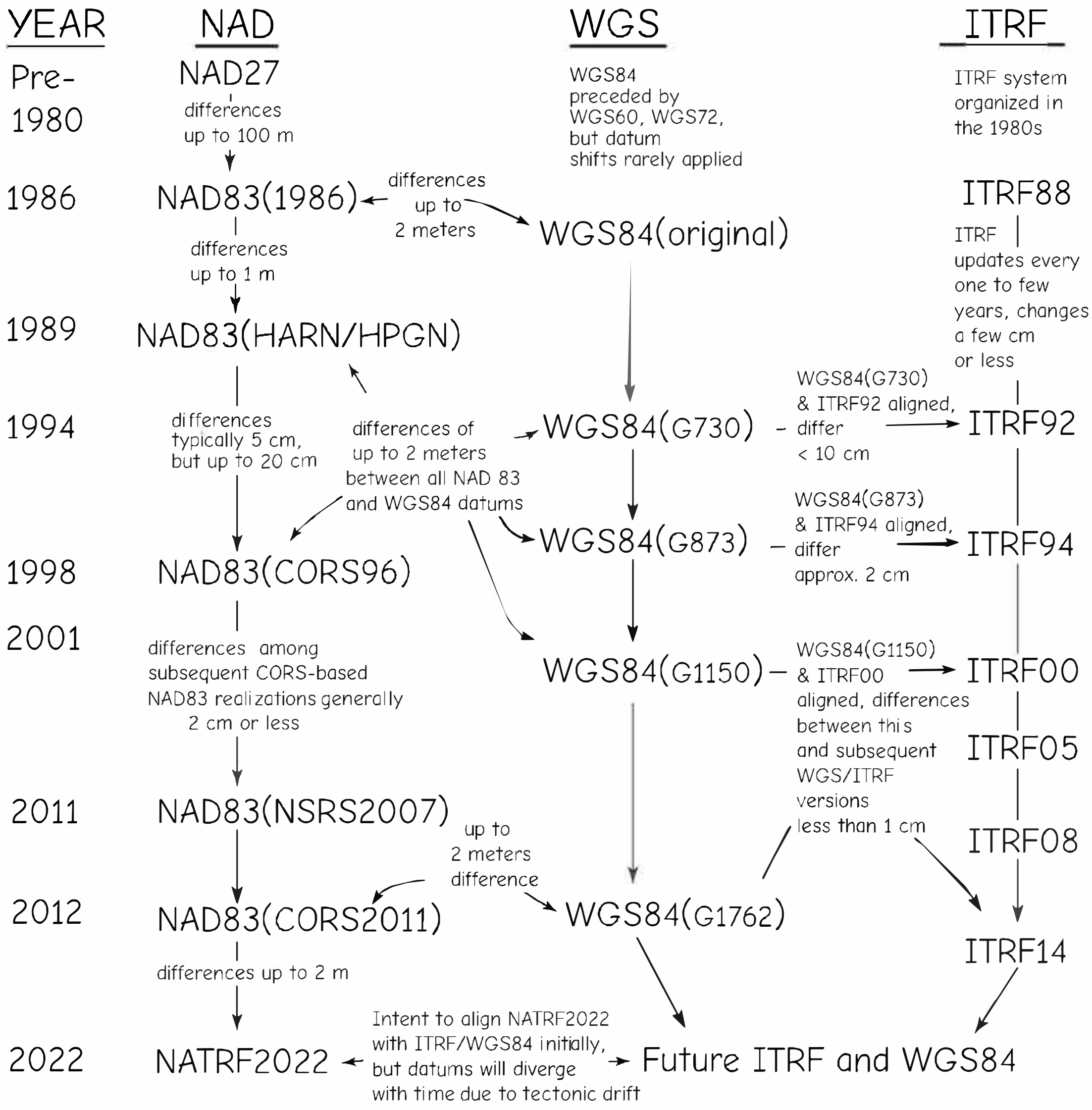


Figure 3-23: This graphic summarizes the evolution of the three main families of datums used in North America. As the datums have been adjusted, horizontal positional differences between survey marks have varied, within the ranges shown. "Aligned" datums (e.g., WGS84(G1150) and ITRF00) may be considered equivalent for most purposes when applying datum transformations.

While locations in the NAD83(xx) and the ITRF/WGS84 datums commonly differ by over a meter, datum shifts internal to these groupings have become small for recent datums. Differences between NAD83(HARN) and NAD83(xx) datums may be up to 20 cm, but are typically less than 4 cm, so these datum realizations may be considered equivalent if accuracy limits are above 20 cm, and perhaps as low as 4 cm. The differences between NAD83(CORS96) and NAD83(2011) are often a few centimeters, as are the differences among ITRF realizations, for example, 91, 94, 00, 05, and 08.

There will be new datum realizations, each requiring additional transformations in the future. The ITRF datums are released every few years, requiring new transformations to existing datums each time. As of this writing, the NGS has released the NAD83(2011) coordinates a nationwide adjustment of passive survey marks and multiyear observations at GNSS/GPS CORS stations.

There is a plan to substantially update the datums used in North America, with the introduction of the *North American Terrestrial Reference Frame of 2022* (NATRF2022). This will initially align official datums for the U.S. with the ITRF and WGS84 datums, removing much of the positional differences for points expressed in these different systems at the time of estimation. It will entail a shift, up to two meters (six feet) in NAD83(2011) coordinates to NATRF2022 coordinates.

Although the NATRF2022 and the ITRF/WGS84 systems will be aligned initially, current plans fix the NATRF2022 to the included tectonic plates, and so will drift from the ITRF positions through time. The transformation will be mathematically simple, using the time since initiation and location, and we expect the US NGS to produce tools to calculate a datum shift given any epoch.

Prior to this decade, differences in datum transformation were usually lower than spatial data error, so it caused few problems. GNSS receivers can now provide centimeter-level accuracy in the field, so what were once considered small datum discrepancies are now apparent. The datum transformation method within any hardware or software system should be documented and the accuracy of the method known before it is adopted. Unfortunately, much data are now degraded because of improper datum transformations.

There are a number of factors that we should keep in mind when applying datum transformations. First, changing a datum changes our best estimate of the coordinate locations of most points. These differences may be small and ignored with little penalty in some specific instances, typically when the changes are smaller than the spatial accuracy required for our analysis. However, many datum shifts are quite large, up to tens of meters. One should know the magnitude of the datum shifts for the area and datum transformations of interest.

Second, datum transformations are estimated relationships that are developed with a specific data set and for a specific area and time. There are spatial errors in the transformations that are specific to the input and datum version. There is no generic transformation between NAD83 and WGS84. Rather, there are transformations between specific versions of each, for example, from NAD83(96) to WGS84(1150).

Finally, GIS projects should not mix datums except under circumstances when the datum shift is small relative to the requirements of the analysis. Unless proven otherwise, all data should be converted to the same coordinate system, based on the same datum. If not, data may misalign.

Vertical Heights and Datums

In its simplest definition, a *vertical datum* is a reference that we use for measuring heights. We commonly specify a vertical datum using a measured, constant gravity (equipotential) surface (Figure 3-24). We then combine these with carefully measured control heights above a specific equipotential surface to define surface heights. As noted in the geoid section on page 92, most government or other organizations use a specific geoid as a reference surface for height, although not everyone adopts the same geoid. Governments adopt "hybrid" geoids that combine their own precise vertical surveys with gravity measurements and models.

Geodesists and surveyors use the term *orthometric heights* to refer to what most of us think of as elevations. This is to clearly refer to our standard heights above our reference surface, different from other height measurements they sometimes use. The orthometric height is the distance from a standard equipotential surface to another level, with the path between the surfaces always at right angle to all intervening gravity surfaces. Orthometric heights have replaced our elevations above mean sea level because, as mentioned earlier, modern vertical heights are referenced to a geoid.

For much of history prior to satellites, *leveling surveys* were used for establishing heights. A standard, seaside bench mark was selected, and distances and elevation differences precisely measured from there to known points. Leveling surveys give the heights of points along their path. Bench marks established at these points were then used to set nearby heights. Early leveling surveys were performed with simple instruments, for example, by *spirit leveling,* using plumb bobs and bubble or tube levels. Horizontal rods were placed between succeeding vertical posts to physically measure height.

The number, accuracy, and extent of leveling surveys increased substantially in the 18th and 19th centuries. Epic surveys that lasted decades were commissioned, such as the Great Arc from southern India to the Himalayas. These surveys were performed at substantial capital and human expense; in one portion of the Great Arc, more than 60% of the field crews died over a six-year period due to illness and mishaps.

Most leveling surveys from the late 1700s through the mid-20th century employed *trigonometric leveling*. This

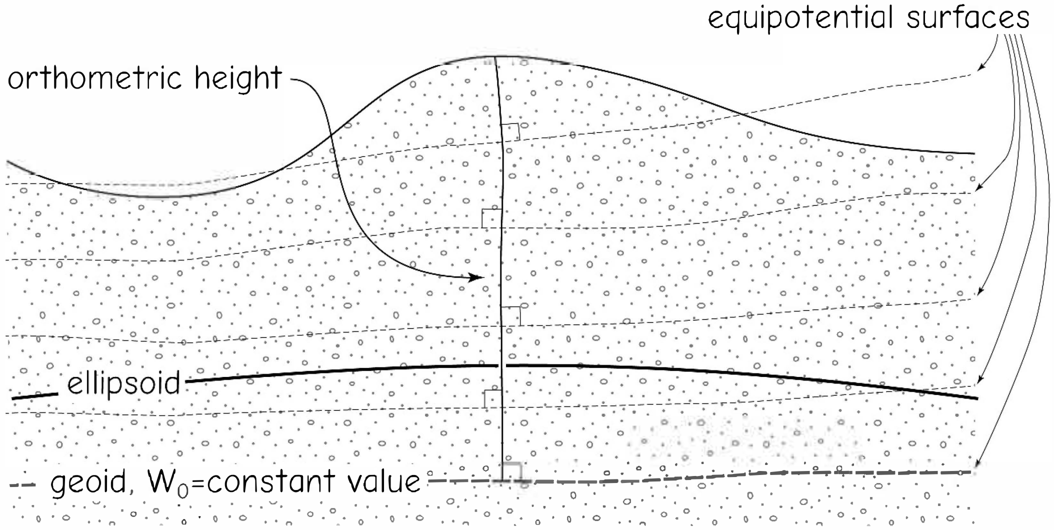


Figure 3-24: Heights in North America are referenced to a geoid, corresponding to a given equipotential surface. All the points on an equipotential surface have the same gravitational pull, and they may be envisioned as layers with decreasing strength at higher levels. Heights are usually specified as orthometric, meaning at right angles to all equipotential surfaces along their path. Because potential surfaces may undulate, orthometric heights may be curved lines, although usually only slightly so.sually only slightly so.

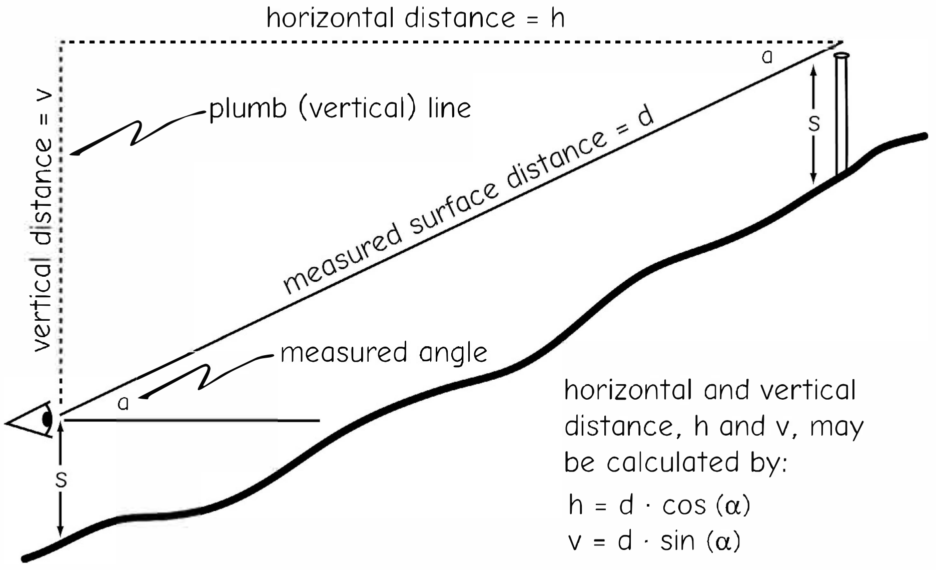


Figure 3-25:Leveling surveys often employ optical measurements of vertical angle (α) with measurements of surface distance (d) and knowledge of trigonometric relationships to calculate horizontal distance (h) and vertical distance (v).

method uses optical instruments and trigonometry to measure changes in height, as shown in Figure 3-25. Surface distance along the slope was measured to avoid the tedious process of establishing vertical posts and leveling rods.The vertical angle was also measured from a known station, typically by a small telescope fitted with a precisely scribed angle gage. Surface distance would then be combined with the measured vertical angle to calculate the horizontal and vertical distances. Early surveys measured surface distance along the slope with ropes, metal chains, and steel tapes. Modern height measurements have largely replaced trigonometric measurements, and primarily use a variety of laser and satellite-based methods.

In North America, we no longer use mean sea-level as a base for orthometric heights (elevations); except for specific projects near the seashore, orthometric heights above a specific vertical datum are now our standard elevation specification. We stopped using mean sea level because it varies too much in time and space. The mean varies in time because of daily through multi-decadal solar and lunar cycles, and across the globe because of persistent differences in water density with temperature, salinity, and ocean currents. Global sea level has been rising over the past century, so the mean at any one seaside station will depend on the length of measurement, even for stations collecting for longer than the 19-year lunar/solar cycles. The mean sea level will differ from Miami to New York, or Amsterdam to Genoa. We weren't able to address this variation until the past few decades, after which methods improved to where the discrepancies in sea levels across the globe became apparent.

Since we want a surface that is consistent in time and space, most countries have picked one or a set of tidal stations, and based orthometric heights relative to a geoidal height passing through or near the station height(s). North American orthometric heights are based on a height specified relative to a long-term tidal gage in Quebec. In mainland Australia, heights are relative to measurements averaged over 30 tidal gages spread along the coast, because they have an approximately 1 meter decline in the geoid height relative to tidal gage measurements from the northeast to the southwestern part of the country. Various European countries adopt base points near different long-term tidal gages, or if landlocked, for points related to gages in adjacent countries. Most countries then adopt an appropriate geoid

and assign a standard orthometric height for the mean gage measurement, and specify all elevations relative to this height.

The geoid adopted for height reference is often a specific gravitational equipotential surface. This is a surface where the pull of gravity is at a specified, constant amount (Figure 3-24). For example, Canada defined the equipotential surface at a gravity value of

$$W_0 = 62,636,856.0 \ m^2s^{-2}$$

as the reference for the Canadian Vertical Datum of 2013 (CGVD2013). Different countries may select different W_0 values, usually corresponding to a single or set of tidal gages on nearby coastlines, but they don't all assign the calculated mean sea level a height of zero. Nonzero heights may be assigned to best match historical data, or when a mean of several stations is used.

Orthometric heights (elevations) in North America are defined as the vertical distance measured from our adopted reference geoid to the ground surface height, along a line that is always at right angles to all intervening equipotential surfaces (Figure 3-24). This height line may bend, as there are often small undulations in the successive equipotential surfaces. The height paths are not the same as a straight line normal to the ellipsoid and up to the surface, and not the same as a straight line that is normal to the geoid surface at the starting point.

Because the zero height may differ among countries, you must be careful when mixing heights across countries. Orthometric heights referenced to one geoid and set of bench marks in Poland may differ from heights referenced to another set in the Netherlands, or ones in Jamaica different from Florida. Unless heights are adjusted, they may be inconsistent when combined across vertical datums. Cooperation among governments is common; for example, datums are compatible across the United States, Canada, and Mexico in North America, and there is a European Vertical Reference System to unify European height datums.

Height datums have varied through time, so care should be taken when combining height data even within any one country. Geoids were fit for North America infrequently before 1990, and several times since. Geoids are named for their target or effective release year, for example, GEOID96 for the North American geoid published in 1996. Geoid versions were subsequently developed for 1999, 2003, 2006, and 2009, with three versions fit for 2012 (an initial, a 2012A, and 2012B). New vertical coordinate data should be developed with reference to the newest datums, as there has been a steady increase in accuracy, coverage, and consistency through time.

The first continental vertical datum in North America was the *National Geodetic Vertical Datum* of 1929, also referred to as NGVD29. Vertical leveling was adjusted to 26 tidal gages, including 5 in Canada, to match measured local mean sea level. Geodesists realized that mean sea level varied across the continent, but assumed these differences would be similar or smaller than measurement errors. They wanted to avoid confusion caused by seaside bench marks having heights that differed from mean sea level.

The latest North American datum is labeled NAVD88. This datum is based on over 600,000 kilometers (373,000 miles) of control leveling performed since 1929, and also reflects geologic crustal movements or subsidence that may have changed bench mark elevation. NAVD88 was fixed relative to only one tidal station because improved measurements yielded errors much smaller than among-station differences in mean sea level, as noted before.

Improved surface, aerial, and satellite gravity measurements, particularly the NASA GRACE and ESA GOCE satellite missions, have led to a dense network of gravity measurements, including regions far from coastal tidal stations. These measurements are combined with previous surveys to update geoid models and allow calculation of the geoidal height at any point on the ellipsoid. Now we most often combine mod-

```
DESIGNATION -  E 58
PID         -  FB1004
STATE/COUNTY-  NC/MADISON
COUNTRY     -  US
USGS QUAD   -  SPRING CREEK (1946)
FB1004                        *CURRENT SURVEY CONTROL
_______________________________________________________________________
NAD 83(2011) POSITION- 35 47 30.13346(N) 082 51 55.76123(W)  ADJUSTED
NAD 83(2011) ELLIP HT-   623.632 (meters)        (06/27/12)  ADJUSTED
NAD 83(2011) EPOCH   -  2010.00
NAVD 88 ORTHO HEIGHT -   653.568 (meters)     2144.25  (feet) ADJUSTED
_______________________________________________________________________
NAD 83(2011) X  -    643,358.550 (meters)                     COMP
NAD 83(2011) Y  - -5,139,947.911 (meters)                     COMP
NAD 83(2011) Z  -  3,709,833.794 (meters)                     COMP
LAPLACE CORR    -          -2.95  (seconds)                   DEFLEC12A
GEOID HEIGHT    -         -29.93  (meters)                    GEOID12A
DYNAMIC HEIGHT  -        652.892 (meters)     2142.03  (feet) COMP
MODELED GRAVITY -    979,578.6   (mgal)                       NAVD 88
```

Figure 3-26: A portion of a data sheet for a vertical control bench mark.

els of geoidal height with measurements of ellipsoidal height (easily given by GNSS systems, described in Chapter 5) to establish orthometric heights.

At this writing, the most current model for North America, GEOID12B, incorporates the best available gravity data with bench marks, leveling, and GPS/GNSS surveys. It has integrated nearly 23,000 vertical bench marks to estimate geoidal and orthometric heights. These heights are known across the continent, and reported on NGS data sheets for vertical bench marks (Figure 3-26). The bench mark sheets also note the vertical datum (here NAVD88), the geoid model (GEOID12), the orthometric height (here 653.568 meters), and the ellipsoidal and geoidal heights. Hybrid vertical datums we use are not entirely independent of horizontal datums, so we should pair our horizontal/vertical datums when combining/converting coordinate data (Figure 3-27).

There is currently an effort to modernize the North American vertical datum, in concert with the horizontal NATRF2022 datum. This will integrate airborne gravity surveys of the entire U.S. and its holdings, to yield a geoid surface estimate accurate to within 1 cm. It will also result in vertical height shifts

NGVD29, no geoid
with
NAD27, NAD83(1986)

NAVD88, GEOID03
with
NAD83(1996)

NAVD88, GEOID09
with
NAD83(NSRS2007)

NAVD88, GEOID12B
with
NAD83(2011)

Figure 3-27: Recommended pairing for horizontal and vertical datums in North America.

of up to 1.3 m (4 feet) from NAVD88 to the new datum.

Because vertical datums differ among regions, and have changed through time within most regions, datum confusion often reigns. Failure to adjust for height differences between vertical datums has caused many errors in height reference, both for older and modern height measurements.

These errors are becoming more commonplace with the widespread use of inexpensive, precise satellite positioning, and with high accuracy laser positioning. Knowledge of the sequence of vertical datums, associated geoid evolution, and vertical datum conversion tools are needed to avoid vertical measurement errors.

Vdatum

Given that vertical datums and associated geoids change through time, the United States National Geodetic Survey (NGS) has created a tool, VDatum, to estimate conversions among vertical datums in the U.S. (Figure 3-28). VDatum calculates the vertical difference from one datum to another at any given horizontal coordinate location and height. Conversions are provided between the 1929 and modern datums, between WGS84/ITRF and NAVD datums, and between various ellipsoid versions within the NAVD88 datum.

Because the vertical datum shift will vary as a function of position, a latitude and longitude must be provided, and because the shift may also depend somewhat on elevation, a vertical height entered. As shown in the example in Figure 3-28, the shifts can be quite large, particularly when converting between NAVD and WGS84/ITRF, and also from NGVD1929 to NAVD88 datums. The vertical datum shift typically changes slowly with distance, so one offset may be suitable for all height shifts over a few to tens of square kilometers. The amount of error and "safe" distance to span varies by region, so the magnitude of the transformation should be verified at several points across any new study area to see how broadly an offset may be used.

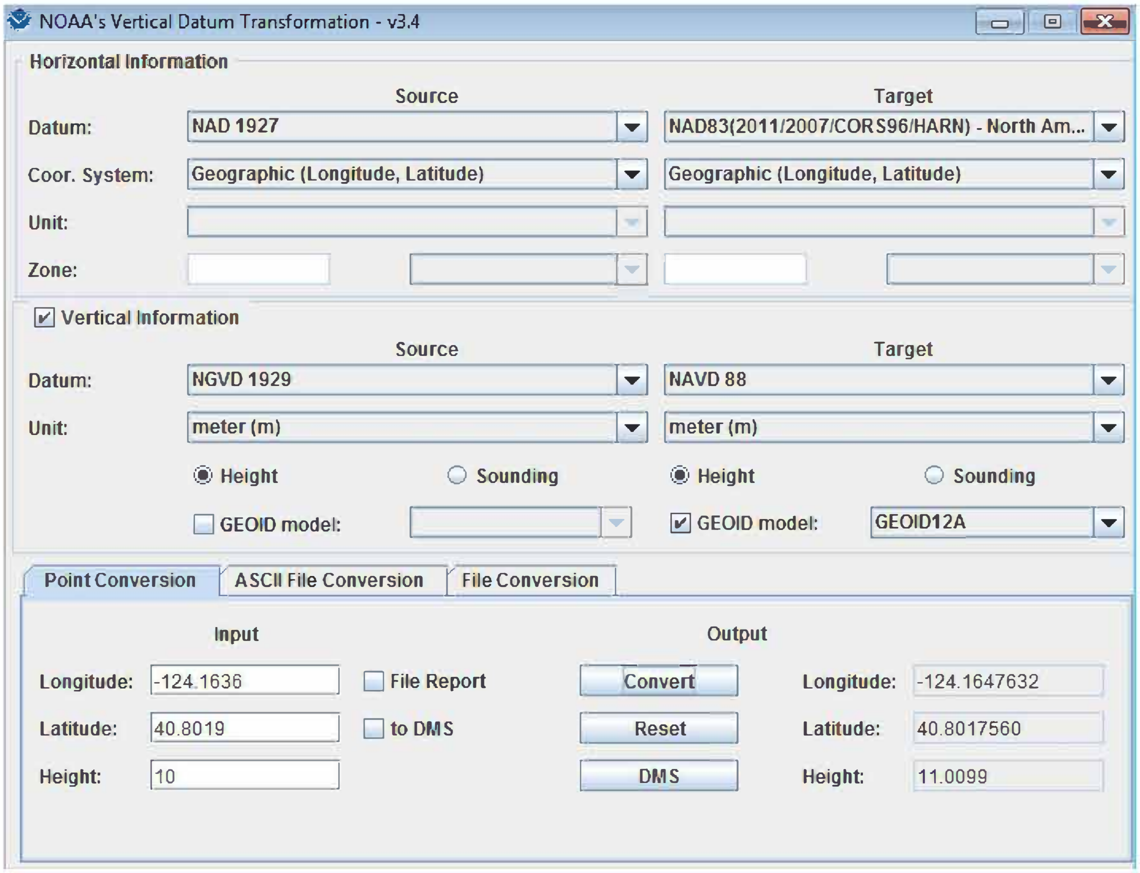


Figure 3-28: An example of the application of the vertical datum transformation software VDatum.

VDatum may also be used to estimate shifts in height among geoid versions. New geoid surfaces have been estimated approximately every three years since 1996 for North America, and heights at any given point will change between geoids. If heights relative to different geoids are to be combined, one set of heights must be adjusted to match the geoid of the other. This is typically achieved by adding an offset calculated from the models included in VDatum.

As an example, I may have two elevation data sets, both in Eureka, California, near a point with latitude 40.8019, longitude -124.1636, and approximately a 10-meter height. One elevation is measured relative to the GEOID96 version of the NAVD88, and the other using the GEOID12A version. I can use VDatum to calculate the vertical height shift due to this difference in geoids; at that coordinate and height, it estimates a 31 cm, or approximately 1 foot, increase in height between these two geoids. This means I would have to add 31 cm to all my 96 heights before combining them with my 12A heights.

Dynamic Heights

We must discuss another kind of height, called a *dynamic height*, because it is important for certain applications. Dynamic heights measure the change in gravitational pull from a given equipotential surface. Dynamic heights are important when interested in water levels and flows across elevations. Points that have the same dynamic heights can be thought of as being at the same water level. Surprisingly, points with the same dynamic heights often have different orthometric heights (Figure 3-29). To be clear, two distinct points at water's edge on a large lake often do not have the same elevations; often, they are different orthometric heights above our reference geoid. Since orthometric heights are our standard for specifying elevation, this means water may indeed flow uphill relative to our standard height measurement, or as confusingly, a lake may have a different elevation on one shore than on the opposite shore.

To understand why water may flow uphill (from lower to higher orthometric heights), it is important to remember how

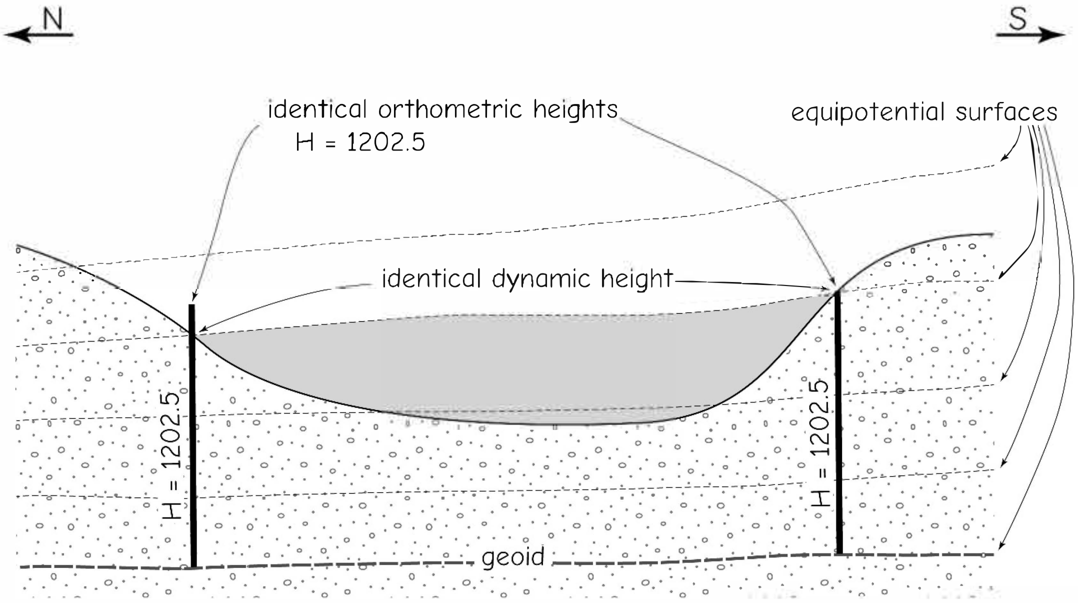


Figure 3-29: An illustration of how dynamic heights and orthometric heights may differ, and how equal orthometric heights may correspond to different heights above the water level on a large lake. Because equipotential surfaces converge, the water level at the northern and southern extremes of a lake will have different orthometric heights. Dynamic heights and water levels are equal across an equipotential surface.

orthometric heights are defined. An orthometric height is the distance, in the direction of gravitational pull, from the geoid up to a point. But remember, the geoid is a specified gravity value, an “equipotential” surface, where the pull of gravity is at some specified level. As we move up from the geoid toward the surface, we pass through other equipotential surfaces, each at a slightly weaker gravitational force, until we arrive at the surface point. But these gravity surfaces are not always parallel, and may be more closely packed in one portion of the globe than another.

There are two key points. First, water spreads out to level across an equipotential surface, absent wind, waves, and other factors. The water level in a still bathtub, pond, or lake has the same equipotential surface at one end as another. Gravity ensures this. Second, the equipotential surfaces are closer together when nearer the mass center of Earth. As the equipotential surfaces converge, or become “denser,” the water surface seems to dip below our fixed orthometric height.

Because water follows an equipotential surface, and because the Earth’s polar radius is less than the equatorial radius, the orthometric heights of the water surface on large lakes are usually different at the north and south ends. For example, as you move farther north in the Northern Hemisphere, the equipotential surfaces converge due to the smaller polar radius, with increased gravitational pull (Figure 3-29). An orthometric height is a fixed height above the geoidal surface, so the northern orthometric height will pass through more equipotential surfaces than the same orthometric height at a more southerly location. An orthometric height of the water surface at the south end of the lake will be higher than at the north end. For example, in Lake Michigan, a large lake in North America, the elevation of the the water surface at the south end is approximately 15 cm higher than the elevation of the water surface at the north end.

Dynamic heights are most often used when we’re interested in relative heights for water levels, particularly over large lakes or connected water bodies. Because equal dynamic heights are at the same water level, we can use them when interested in accurately representing hydrologic drop, head, pressure, and other variables related to water levels across distances. But these differences could be confusing when observing bench mark or sea level heights, and underscore again that our height reference is not mean sea level, but rather an estimated geoidal surface.

Local Sea Level Datums

Water height measurements along the U.S. coast are typically reference to local sea level datums. As noted earlier, mean sea level is not zero for almost all points along North America’s coastline. Elevations are measured relative to a geoid. Zero elevation coincides with zero mean sea level at only one standard coastal station in Canada, near the center of the continent. Variations in gravity, currents, salinity, tides, and wind produce mean sea levels that are different from zero by up to several meters (10s of feet) around the rest of the continental rim. But we still need to know the ocean level along the coastline for many practical purposes, including construction, flood protection, and water management. We have established a network of long-term, reference measurement stations along the coastline. We precisely measure both sea level and the station orthometric height, so that we

can tie our standard elevation to local water heights.

Data for measured tidal stations are available from the NOAA web page:

tidesandcurrents.noaa.gov/stations.html

These sites report mean sea level, as well as mean high, low, and extreme water levels (Figure 3-30). Most importantly, they also report the NAVD88 orthometric heights for each tidal station, allowing a conversion from local sea level heights to measured surface elevation.

Figure 3-30 shows data for a station in Seattle measured since 1899. Mean sea level has a local reference height of 6.64 feet, meaning the sea level averages that height above the long-term measurement of a given low water height. The NAVD88 height at the same point is 2.34 feet, which yields a se level height of 6.64 - 2.34, or 4.3 feet. As strange as it may seem at first, the mean sea level at this Seattle station has an elevation of 4.3 feet. Any point nearby that has an elevation less than 4.3 feet will be below sea level, and will likely flood frequently if there is access to the sea. Local construction, water level measurements, or other activities dependent on sea level will reference this station measurements, and the 4.3 foot offset between mean sea level and seaside orthometric heights.

This mean sea level offset varies by location, for example, Port San Luis, CA, has a vertical offset of 2.7 feet, and Vaca Key, FL, has an offset of -0.8 feet. When heights of sea level are important for an analysis, projects should reference the nearest local datum.

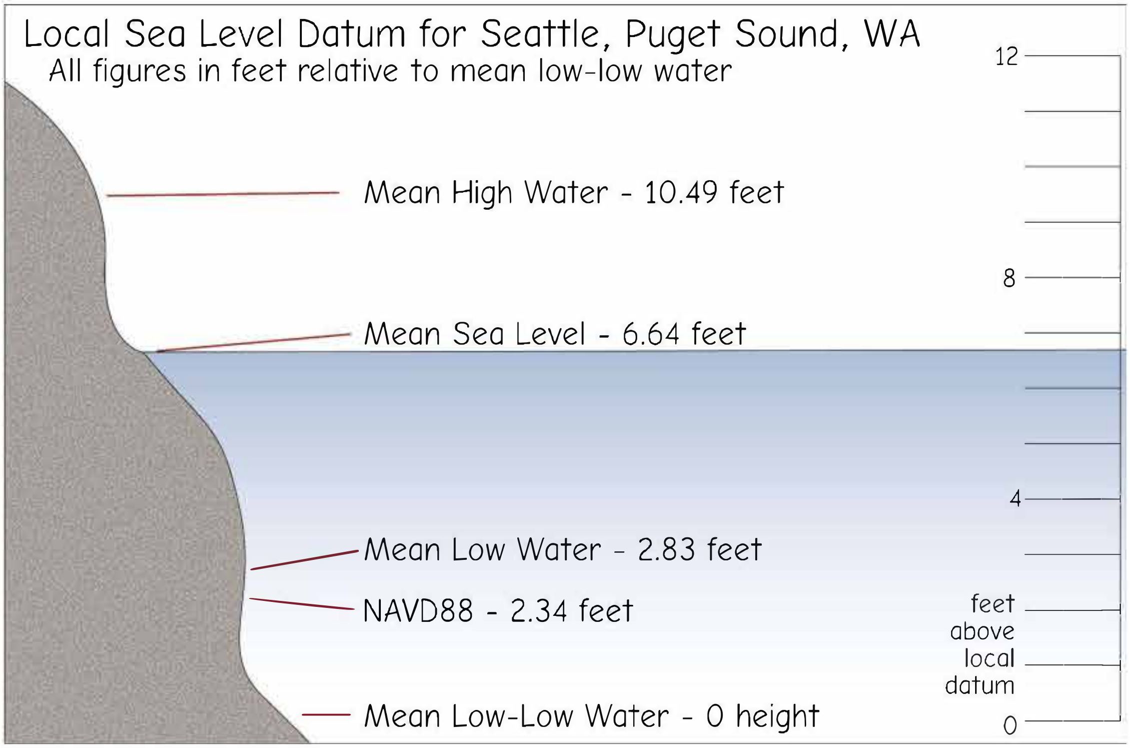


Figure 3-30:An illustration of mean sea level and other measures at a NOAA long-term tidal gage. Note that the mean sea level at this station is 6.64 - 2.34, or 4.3 feet above the NAVD88 zero height.

Map Projections and Coordinate Systems

Datums tell us the latitudes and longitudes of features on an ellipsoid. We need to transfer these from the curved ellipsoid to a flat map. A *map projection* is a systematic rendering of locations from the curved Earth surface onto a flat map surface.

Nearly all projections are applied via exact or iterated mathematical formulas that convert between geographic latitude/longitude pairs and projected X/Y (easting and northing) coordinates. Figure 3-31 shows one of the simpler projection equations, between Mercator and geographic coordinates, assuming a spherical Earth. These equations would be applied for every point, vertex, node, or grid cell in a data set, converting the vector or raster data feature by feature from geographic to Mercator coordinates.

Conversion from geographic (lon, lat) to projected coordinates

Given longitude = λ, latitude = ϕ (all angles in radians)

Mercator projection coordinates are:

$$x = R \cdot (\lambda - \lambda_o)$$
$$y = R \cdot \ln(\tan(\pi/4 + \phi/2))$$

where R is the radius of the sphere at map scale (e.g., Earth's radius), ln is the natural log function, and λ_o is the longitudinal origin (Greenwich meridian)

Inverse equation, from x, y to λ, ϕ:

$$\lambda = x/R + \lambda_o$$
$$\phi = (\pi/2) - 2 \cdot \tan^{-1}[e^{-y/R}]$$

Figure 3-31: Formulas are known for most projections that provide exact projected coordinates, if the latitudes and longitudes are known. This example shows the formulas defining the Mercator projection for a sphere.

Notice that there are parameters we must specify for this projection – here R, the Earth's radius, and λ_o, the longitudinal origin. Different values for these parameters give different values for the coordinates, so even though we may have the same kind of projection (transverse Mercator), we have different versions each time we specify different parameters.

Projection equations must also be specified in the "backward" direction, from projected coordinates to geographic coordinates, if they are to be useful. The projection coordinates in this backward, or "inverse," direction are often much more complicated than the forward direction, but are specified for every commonly used projection.

Most projection equations are much more complicated than the transverse Mercator, in part because most adopt an ellipsoidal Earth, and because the projections are onto curved surfaces rather than a plane. Thankfully, projection equations have long been standardized, documented, and made widely available through proven programming libraries and projection calculators.

Note that each projection defines a Cartesian coordinate system and hence creates *grid north*, a third version of the northern direction, in addition to geographic and magnetic norths. Grid north is the direction of the Y axis in a map projection, and often equals or nearly equals the direction of a meridian near the center of the projected area. Grid north is typically different from geographic and magnetic north for most of the projected region.

Most map projections may be viewed as sending rays of light from a projection source through the ellipsoid and onto a map surface (Figure 3-32). In some projections,

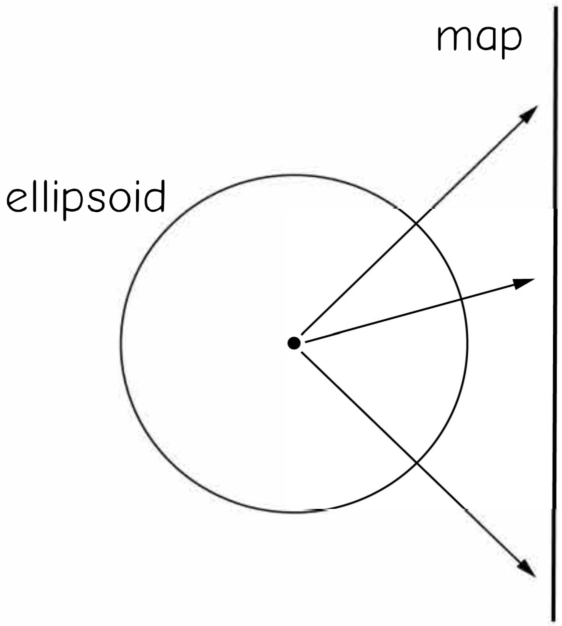


Figure 3-32: A conceptual view of a map projection.

the source is not a single point; however, the basic process involves the systematic transfer of points from the curved ellipsoidal surface to a flat map surface.

Distortions are unavoidable when making flat maps because of the transition from a complexly curved Earth surface to a flat or simply curved map surface. Portions of the rendered Earth surface must be compressed or stretched to fit onto the map. This is illustrated in Figure 3-33, a side view of a projection from an ellipsoid onto a plane.The map surface intersects the Earth at two locations, I_1 and I_2. Points toward the edge of the map surface, such as D and E, are stretched apart. The scaled map distance between d and e is greater than the distance from D to E measured on the surface of the Earth. More simply put, the distance along the map plane is greater than the corresponding distance along the curved Earth surface. Conversely, points such as A and B that lie in between I_1 and I_2 would appear compressed together. The scaled map distance from a to b would be less than the surface measured distance from A to B. Distortions at I_1 and I_2 are zero.

Figure 3-33 demonstrates a few important facts. First, distortion may differ in sense across the map. Parts of the map may have compressed areas or distances relative to the scaled Earth's surface measurements, while other parts may have expanded areas or distances. Second, there are often a few points or lines where distortions are zero and where length, direction, or some other geometric property is preserved. Finally, distortion is usually small near the points or lines of intersection, and increases with increasing distance from the points or lines of intersection.

Different map projections may distort the globe in different ways. The projection source, represented by the point at the middle of the circle in Figure 3-33, may change locations. We may project on to different shapes, and we may place the projection surface at different locations at or near the globe. If we change any of these three factors, we will change how or where our map is distorted. The type and amount of projection distortion may guide the selection of the appropriate projection or limit the area projected.

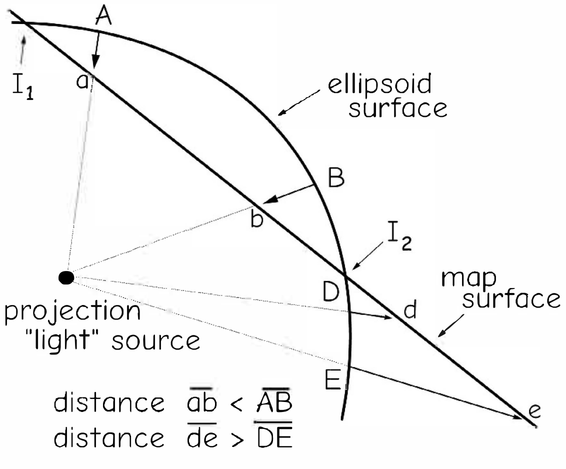


Figure 3-33: Distortion during map projection. This side view shows both expansion and compression of areas on a planar map.

Figure 3-34 shows an example of distortion with a projection onto a planar surface, but from above rather than the side view in Figure 3-33. This planar surface intersects the globe at a line of true scale, the solid circle shown in Figure 3-34. Distortion increases away from the line of true scale, with features inside the circle compressed or reduced in size,while features outside the standard circle are expanded. Calculations show a scale error of -1% near the center of the circle, and increasing scale error in concentric bands outside the circle to over 2% near the outer edges of the projected area.

An approximation of the distance distortion may be obtained for any projection by comparing grid coordinate distances to *great circle distances*. A great circle distance is defined on the surface of the spheroid or ellipsoid (Figure 3-35). The circle distance is the shortest path between two points on the surface of the ellipsoid, and by approximation, Earth.

Figure 3-35 illustrates the calculation of both the great circle and projection, or Cartesian distances for two points in the southern U.S., using the spherical approximation formula introduced in Chapter 2. We use a spherical approximation of the Earth's shape because it is accurate enough for illustration. The difference between this simpler spheroidal method (equal polar and equatorial radii) and an ellipsoidal method is typically much less than 0.1%, and always less than 0.3%, so typically less than 50 cm (1.5 feet) in our example.

Projected (Cartesian) coordinates in this example are in the UTM Zone 15N coordinate system, and derived from the appropriate coordinate transformation equations.

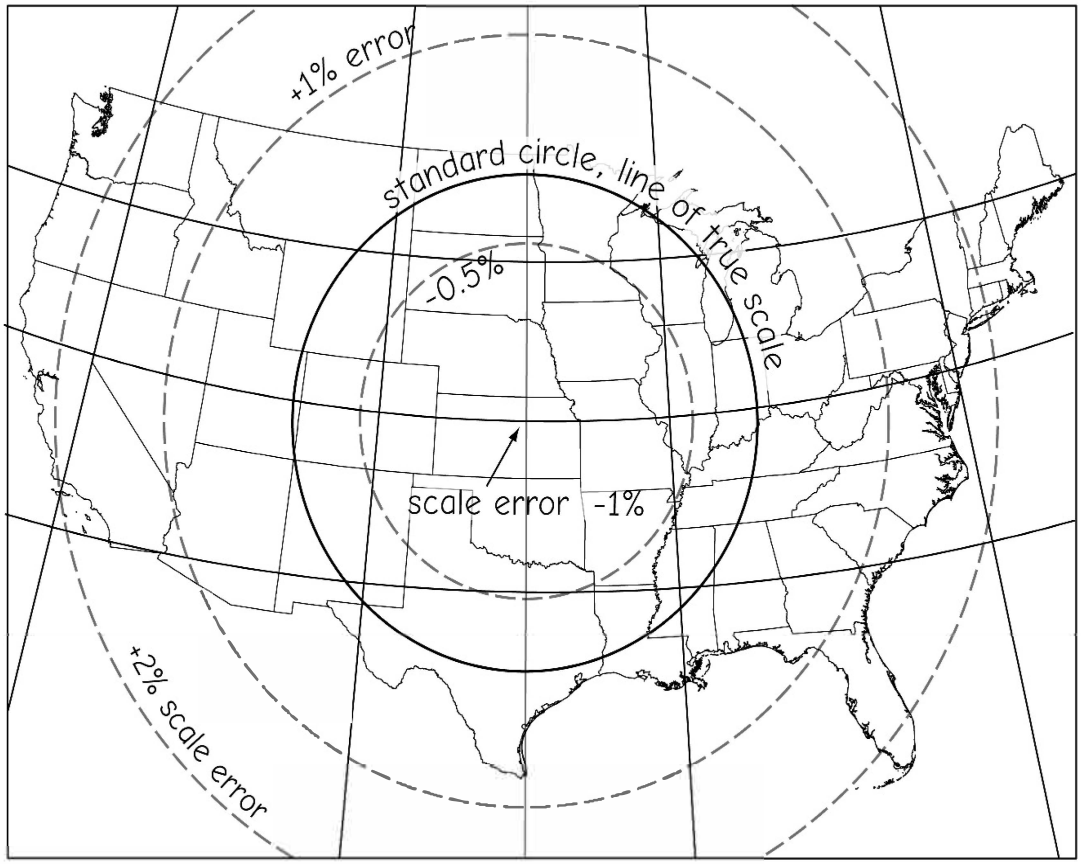


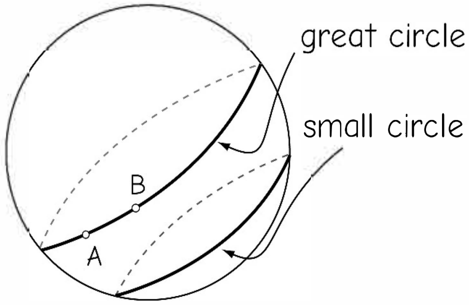
Figure 3-34: Approximate error due to projection distortion for a specific oblique stereographic projection. A plane intersects the globe at a standard circle. This standard circle defines a line of true scale, where there is no distance distortion. Distortion increases away from this line, and varies from -1% to over 2% in this example (adapted from Snyder, 1987).

Great Circle vs. Projected Distance

Spherical Approximation

Using the great circle formula from our example in Chapter 2,

A with latitude, longitude of (ϕ_A, λ_A), and
B, with latitude, longitude of (ϕ_B, λ_B)



The great circle distance from point A to point B is given by the formula:

A corresponding to Baton Rouge, LA = 30.4877456°, -91.1693348°
B corresponding to Houston, Texas = 29.7507171°, -95.370003°

$$d = 6378 \cdot 2\sqrt{\sin^{-1}[(\sin^2(0.368514)) + \cos(30.4877456)\cdot\cos(29.7507171)\cdot\sin^2(2.1003341)]}$$
$$= 412.681 \text{ km}$$

Grid distance (UTM Zone 15N coordinates):
Grid coordinates of Baton Rouge, LA = 675,708.2, 3,374,258.0
Grid coordinates of Houston, Texas = 270,816.1, 3,293,516.3

$$dg = [(X_A - X_B)^2 + (Y_A - Y_B)^2]^{0.5}$$
$$= [(675{,}708.2 - 270{,}816.1)^2 + (3{,}374{,}258.0 - 3{,}293{,}516.3)^2]^{0.5}$$
$$= 412.864 \text{ km}$$

distortion is 412.681 - 412.864 = -0.183 km, or a 183 meter lengthening

Figure 3-35: Example calculation of the distance distortion due to a map projection. The great circle and grid distances are compared for two points on the Earth's surface, the first measuring along the curved surface, the second on the projected surface. The difference in these two measures is the distance distortion due to the map projection. Calculations of the great circle distances are approximate, due to the assumption of a spheroidal rather than ellipsoidal Earth, but are at worst within 0.3% of the true value along the ellipsoid. Note that various great circle distance calculators are available via the World Wide Web, and these often don't specify the formula or Earth radius values used, so different great circle distances may be provided.

Armed with the coordinates for both pairs of points in both the geographic and projected coordinates, we can calculate the distance in the two systems, and subtract to find the length distortion due to projecting from the spherical surface to a flat surface.

Note that web-based or other software may use the ellipsoidal approximation, and may not specify the Earth radii used, so it is best to calculate the values from the original formulas when answers differ substantially.

A straight line between two points shown on a projected map is usually not a straight line nor the shortest path when traveling on the surface of the Earth. Conversely, the shortest distance between points on the Earth surface is likely to appear as a curved line on a projected map. The distortion is imperceptible for large scale maps and over short distances, but exists for most lines.

Figure 3-36 illustrates straight line distortion. This figure shows the shortest distance path (the great circle) between Seattle, USA, and Paris, France. Paris lies almost due east of Seattle, but the shortest path traces a route north of an east-west line. This shortest path is distorted and appears curved by the Plate Carrée projection commonly used for global maps.

Figure 3-36: Curved representations of straight lines are a manifestation of projection distortion. A great circle path, shown above, is the shortest route when flying from Paris to Seattle, and commonly appears distorted when displayed.

Projections may also substantially distort the shape and area of polygons. Figure 3-37 shows various projections for Greenland, from an approximately "unprojected" view from space through geographic coordinates cast on a plane, to Mercator and transverse Mercator projections. Note the changes in size and shape of the polygon depicting Greenland.

Most map projections are based on a *developable surface*, a geometric shape onto which the Earth's surface is projected. Cones, cylinders, and planes are the most common developable surfaces. A plane is already flat, and cones and cylinders may be mathematically "cut" and "unrolled" to develop a flat surface (Figure 3-38). Projections may be characterized according to the shape of the developable surface, as *conic* (cone), *cylindrical* (cylinder), and *azimuthal* (plane). The orientation of the developable surface may also change among projections; for example, the axis of a cylinder may coincide with the poles (equatorial) or the axis may pass through the Equator (transverse).

Note that while the most common map projections used for spatial data in a GIS are based on a developable surface, many map projections are not. Projections with names such as pseudocylindrical, Mollweide, sinusoidal, and Goode homolosine are examples. These projections often specify a direct mathematical projection from an ellipsoid onto a flat surface. They use mathematical forms not related to cones, cylinders, planes, or other three-dimensional figures, and may change the projection surface for different parts of the globe, but generally are used only for display, and not for spatial analysis, because the coordinate systems are not strictly Cartesian.

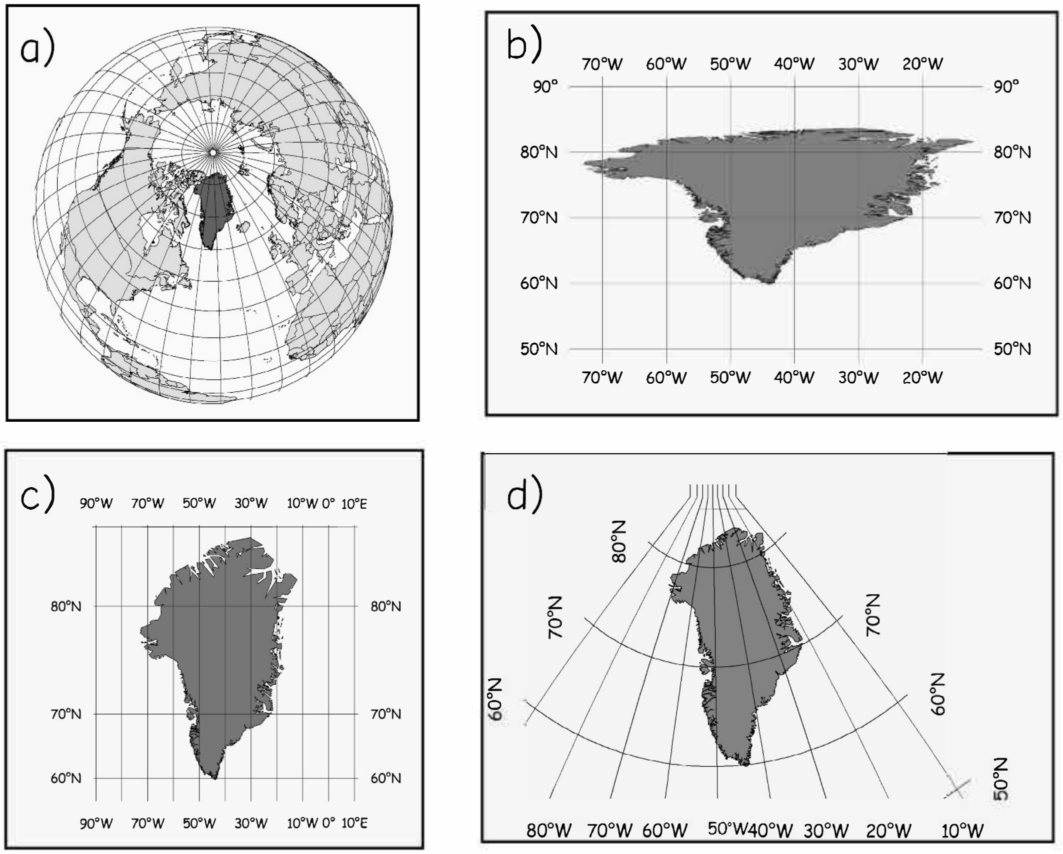


Figure 3-37: Map projections can distort the shape and area of features, as illustrated with these various projections of Greenland, from a) approximately unprojected, b) geographic coordinates on a plane, c) a Mercator projection, and d) a transverse Mercator projection.

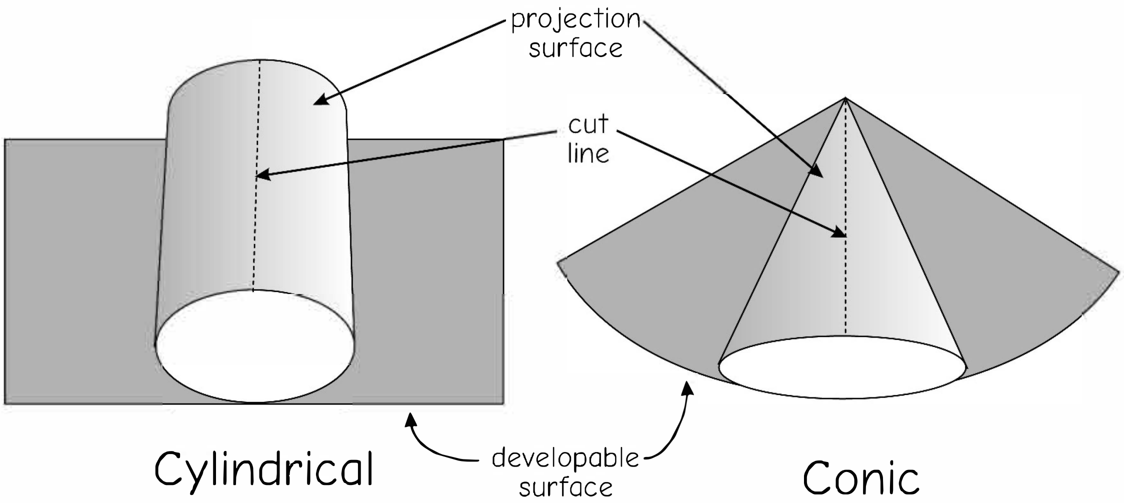


Figure 3-38: Projection surfaces are derived from curved "developable" surfaces that may be mathematically "unrolled" to a flat surface.

Common Map Projections in GIS

There are hundreds of map projections used throughout the world; however, most spatial data in GIS are specified using a relatively small number of projection types.

The Lambert conformal conic and the transverse Mercator are among the most common projection types used for spatial data in North America, and much of the world (Figure 3-39). Standard sets of projections have been established from these two basic types. The Lambert conformal conic (LCC) projection may be conceptualized as a cone intersecting the surface of the Earth, with points on the Earth's surface projected onto the cone. The cone in the Lambert conformal conic intersects the ellipsoid along two arcs, typically parallels of latitude, as shown in Figure 3-39 (top left). These lines of intersection are known as *standard parallels.*

Distortion in a Lambert conformal conic projection is typically smallest near the standard parallels, where the developable surface intersects Earth. Distortion increases in a complex fashion as distance from these parallels increases. This characteristic is illustrated at the top right and bottom of Figure 3-39. Circles of a constant 5-degree radius are drawn on the projected surface at the top right, and approximate lines of constant distortion and a line of true scale are shown in Figure 3-39, bottom. Distortion decreases toward the standard parallels, and increases away from these lines. Distortions can be quite severe, as illustrated by the apparent expansion of southern South America.

Note that sets of circles in an east-west row are distorted in the Lambert conformal conic projection (Figure 3-39, top right). Those circles that fall between the standard parallels typically exhibit a lower distortion than those in other portions of the projected map. This property of a low-distortion band running in an east-west direction between the standard parallels makes the Lambert conformal conic projection popular for mapping areas that are larger in an east-west than a north-south direction. We add little distortion when extending the mapped area in the east-west direction.

Distortion is controlled by the placement of the standard parallels, the lines where the cone intersects the globe. The example in Figure 3-39 shows parallels placed such that there is a maximum distortion of approximately 1% midway between the standard parallels. We reduce this distortion by moving the parallels closer together, but at the expense of increasing distortion outside the zone between the lines.

The transverse Mercator is another common map projection. This map projection may be conceptualized as enveloping the Earth in a horizontal cylinder, and projecting the Earth's surface onto the cylinder (Figure 3-40). The cylinder in the transverse Mercator commonly intersects the Earth ellipsoid along a single north–south tangent, or along two *secant* lines, noted as the lines of true scale in Figure 3-40. A line parallel to and midway between the secants is often called the *central meridian*. The central meridian extends north and south through transverse Mercator projections.

As with the Lambert conformal conic, the transverse Mercator projection has a band of low distortion, but this band runs in a north-south direction. Distortion is least near the line(s) of intersection. The graph at the top right of Figure 3-40 shows a transverse Mercator projection with the central meridian (line of intersection) at W96^{o}. Distortion increases markedly with distance east or west away from the intersection line; for example, the shape of South America is severely distorted in the top right of Figure 3-40. The drawing at the bottom of this same figure shows lines estimating approximately equal scale distortion for a transverse Mercator projection centered on the USA. Notice that the distortion increases as distance from the two lines of intersection increases. Scale distortion error may be maintained below any threshold by ensuring the mapped area is close to these two secant lines intersecting the globe. Transverse Mercator projections are often used for areas that extend in a

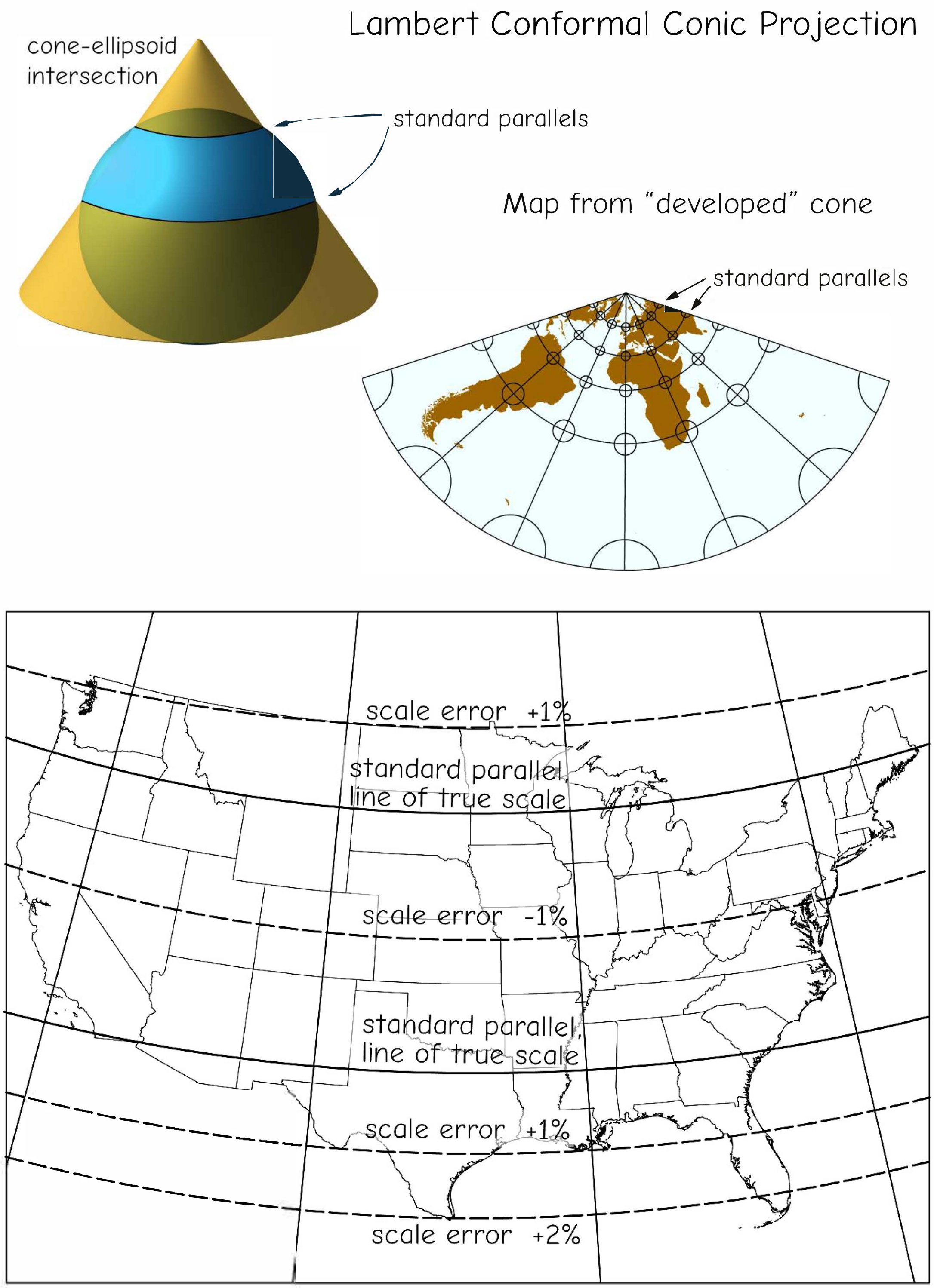
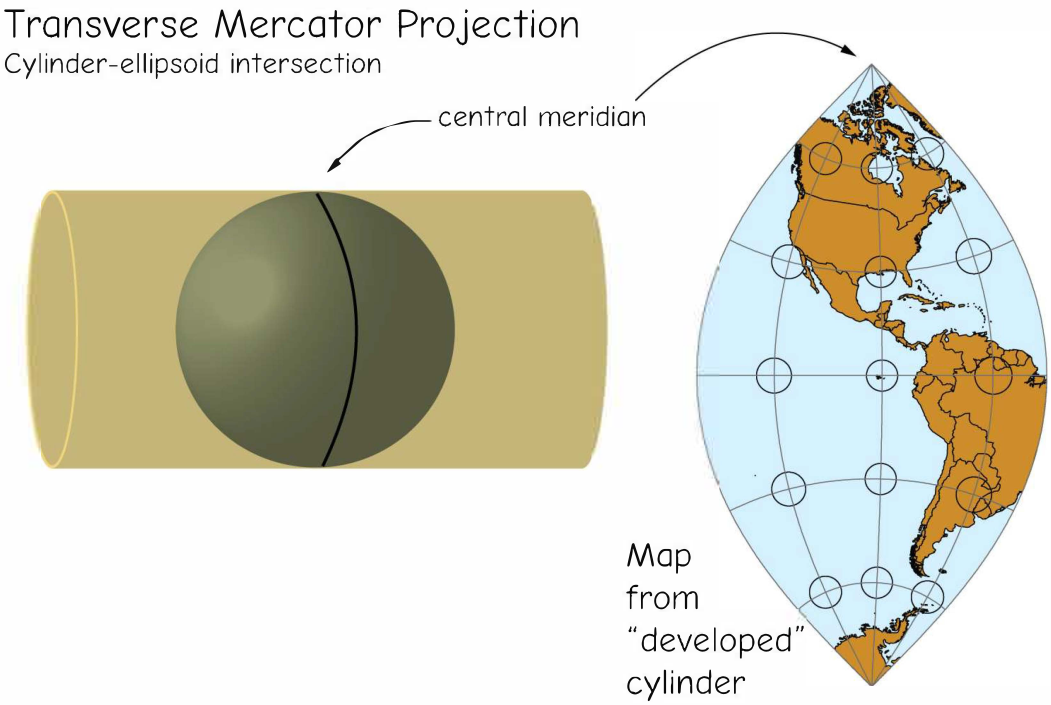


Figure 3-39: Lambert conformal conic (LCC) projection (top) and an illustration of the scale distortion associated with the projection. The LCC is derived from a cone intersecting the ellipsoid along two standard parallels (top left). The "developed" map surface is mathematically unrolled from the cone (top right). Distortion is primarily in the north-south direction, and is illustrated in the developed surfaces by the deformation of the 5-degree diameter geographic circles (top) and by the lines of approximately equal distortion (bottom). Note that there is no scale distortion where the standard parallels intersect the globe, at the lines of true scale (bottom, adapted from Snyder, 1987).



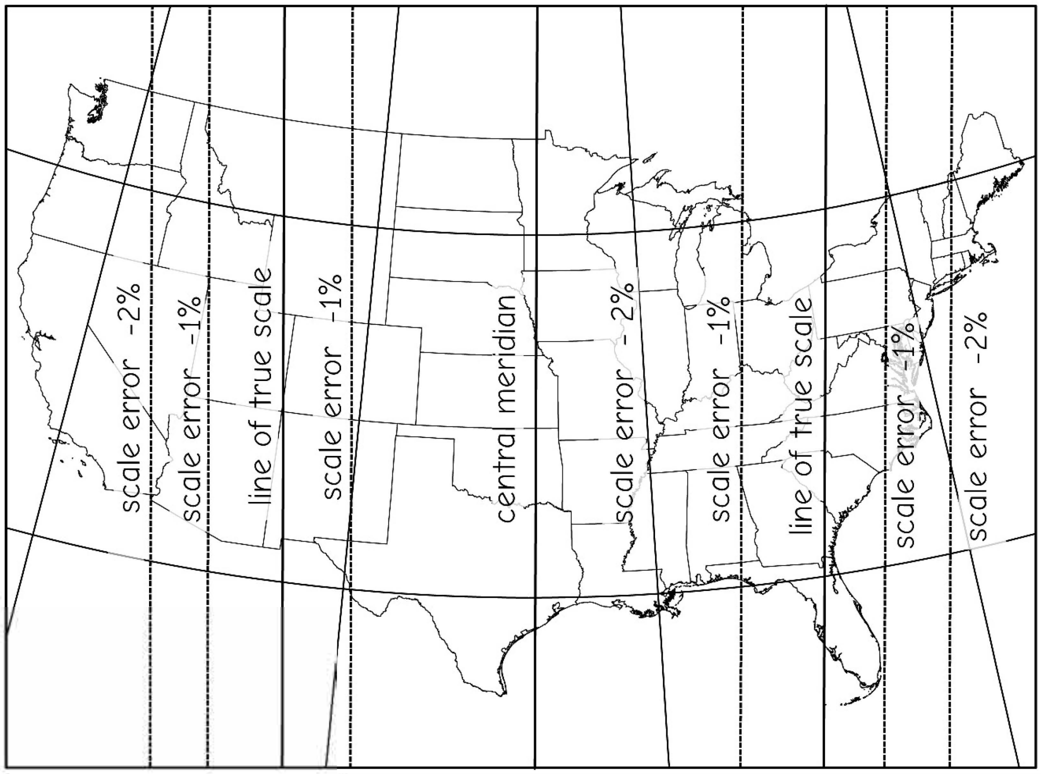


Figure 3-40: Transverse Mercator (TM) projection (top), and an illustration of the scale distortion associated with the projection (bottom). The TM projection distorts distances in an east-west direction, but has relatively little distortion in a north-south direction. This TM intersects the sphere along two lines, and distortion increases with distance from these lines (bottom, adapted from Snyder, 1987).

north-south direction, as there is little added distortion extending in that direction.

Different projection parameters may be used to specify an appropriate coordinate system for a region of interest. Specific standard parallels or central meridians are chosen to minimize distortion over a mapping area. An origin location, measurement units, x and y (or northing and easting) offsets, a scale factor, and other parameters may also be required to define a specific projection.

The State Plane Coordinate System

The State Plane Coordinate System (SPCS) is a standard set of projections for the United States. The SPCS specifies positions in Cartesian coordinate systems for each state. There are one or more zones in most states, with slightly different projection parameters in each State Plane zone (Figure 3-41). Multiple State Plane zones are used to limit distortion errors due to map projections.

State Plane systems ease surveying, mapping, and spatial data development in a GIS, particularly when whole counties or larger areas are covered. The State Plane system provides a common coordinate reference for horizontal coordinates over county to multi-county areas while limiting distortion error to specified maximum values. Most states have adopted zones such that projection distortions are kept below one part in 10,000. Some states allow larger distortions (e.g., Montana, Nebraska) for the sake of having only one state plane zone. SPCSs are used in many types of work, including property surveys, property subdivisions, large-scale construction projects, and photogrammetric mapping, and the zones and SPCSs are often adopted for GIS.

One State Plane projection zone may suffice for small states. Larger states commonly require several zones, each with a different projection, for each of several geographic zones of the state. For example, Delaware has one State Plane coordinate zone, while California has six, and Alaska has 10 State Plane coordinate zones, each corresponding to a different projection within the state. Zones are added to a state to ensure acceptable projection distortion within all zones (Figure 3-42, left). Zone boundaries are defined by county, parish, or other municipal boundaries. For example, the Minnesota south/central zone boundary runs approximately east–west through the

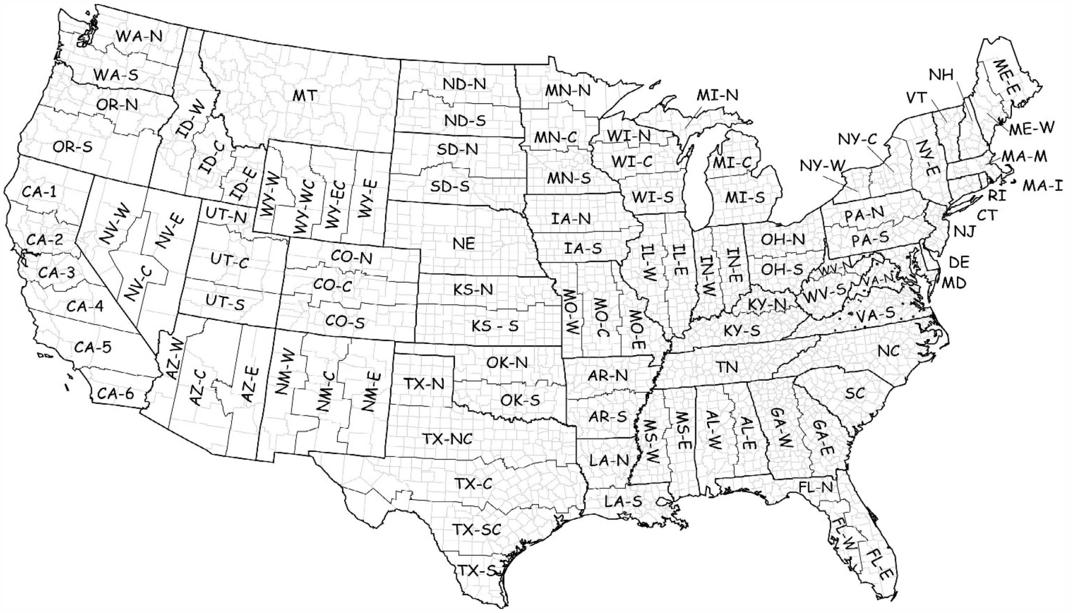


Figure 3-41: State plane zone boundaries, NAD83.

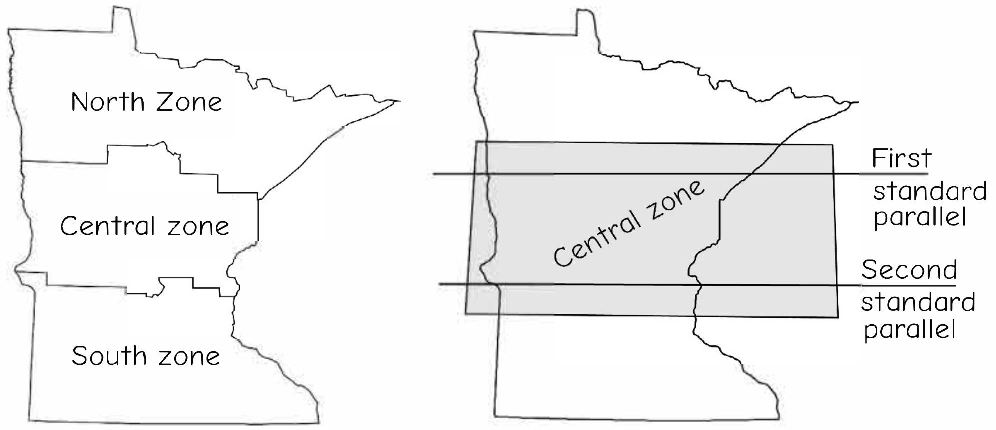


Figure 3-42: The State Plane zones of Minnesota, and details of the standard parallel placement for the Minnesota central State Plane zone.

state along defined county boundaries (Figure 3-42, left).

Most State Plane coordinate systems are based on one of two types of map projections: the Lambert conformal conic or the transverse Mercator projections. Because distortion in a transverse Mercator increases with distance from the central meridian, this projection type is most often used with states or zones that have a long north-south axis (e.g., Illinois or New Hampshire). Conversely, a Lambert conformal conic projection is most often used when the long axis of a state or zone is in the east-west direction (examples are North Carolina and Virginia).

Standard parallels for the Lambert conformal conic projection, described earlier, are specified for each State Plane zone. These parallels are placed at one-sixth of the zone width from the north and south limits of the zone (Figure 3-42, right). A zone central meridian is specified at a longitude near the zone center. This central meridian points at grid north; however, all other meridians converge to this central meridian, so they do not point to grid north. The Lambert conformal conic is used for State Plane zones for 31 states.

As noted earlier, the transverse Mercator specifies a central meridian. This central meridian defines grid north in the projection. A line along the central meridian points to geographic and grid north, and specifies the Cartesian grid direction for the map projection. All parallels of latitude and all meridians except the central meridian are curved for a transverse Mercator projection, and hence these lines do not parallel the grid x or y directions. The transverse Mercator is used for 22 State Plane systems (the sum of states is greater than 50 because both the transverse Mercator and Lambert conformal conic are used in some states, e.g., Florida).

Finally, note that more than one version of the State Plane coordinate system has been defined. Changes were introduced with the adoption of the North American Datum of 1983. Prior to 1983, the State Plane projections were based on NAD27. Changes were minor in some cases, and major in others, depending on the state and State Plane zone. Some states, such as South Carolina, Nebraska, and California, dropped zones between the NAD27 and NAD83 versions (Figure 3-43). Others maintained the same number of State Plane zones, but changed the projection by the placement of the meridians, or by switching to a metric coordinate system rather than one using feet, or by shifting the projection origin. State Plane zones are sometimes identified by the Federal Information Processing System (FIPS) codes, and most codes are similar across NAD27 and NAD83 versions. Care must be taken when using legacy data to identify the version of the State Plane coordinate system used because the FIPS and State Plane zone

designators may be the same, but the projection parameters may have changed from NAD27 to NAD83.

Conversion among State Plane projections may be further confused by the various definitions used to translate from feet to meters. The metric system was first developed during the French Revolution in the late 1700s, and it was adopted as the official unit of distance in the United States, by the initiative of Thomas Jefferson. President Jefferson was a proponent of the metric system because it improved scientific measurements, was based on well-defined, integrated units, reduced commercial fraud, and improved trade within the new nation. The conversion was defined in the United States as one meter equal to exactly 39.97 inches. This yields a conversion for a *U.S. survey foot* of:

1 foot = 0.3048006096012 meters

Unfortunately, revolutionary tumult, national competition, and scientific differences led to the eventual adoption of a different conversion factor in Europe and most of the rest of the world. They adopted an *international foot* of:

1 foot = 0.3048 meters

The U.S. definition of a foot is slightly longer than the European definition, by about one part in five million. Both conversions are used in the U.S., and the international conversion elsewhere. The European conversion was adopted as the standard for all measures under an international agreement in the 1950s. However, there was a long history of the use of the U.S. conversion in U.S. geodetic and land surveys. Therefore, the U.S. conversion was called the U.S. survey foot. This slightly longer metric-to-foot conversion factor should be used as the default for conversions among geodetic coordinate systems within the United States, for example, when converting from a State Plane coordinate system specified in feet to one specified in meters.

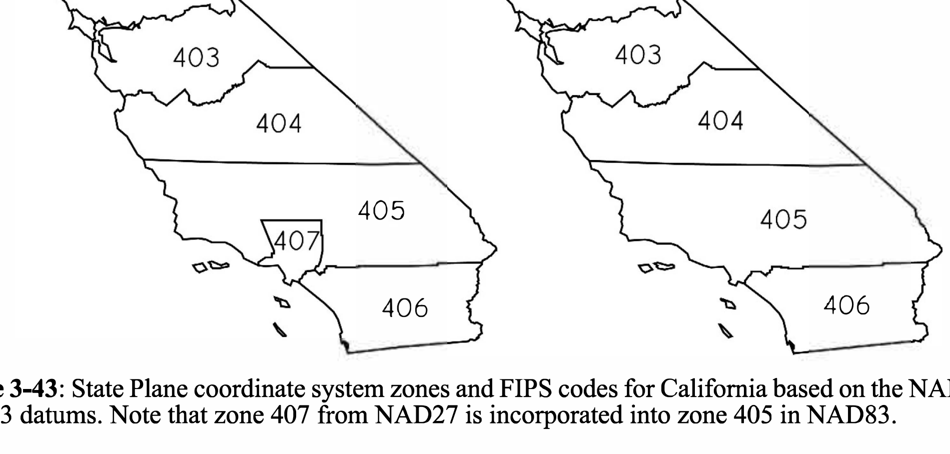


Figure 3-43: State Plane coordinate system zones and FIPS codes for California based on the NAD27 and NAD83 datums. Note that zone 407 from NAD27 is incorporated into zone 405 in NAD83.

Universal Transverse Mercator Coordinate System

The Universal Transverse Mercator (UTM) coordinate system is another standard, distinct from the State Plane system. The UTM is a global coordinate system, based on the transverse Mercator projection. It is widely used in the United States and other parts of North America, and is also used in many other countries.

The UTM system divides the Earth into zones that are 6 degrees wide in longitude and extend from 80 degrees south latitude to 84 degrees north latitude. UTM zones are numbered from 1 to 60 in an easterly direction, starting at longitude 180 degrees West (Figure 3-44). Zones are further split north and south of the Equator. Therefore, the zone containing most of England is identified as UTM Zone 30 North, while the zones containing most of New Zealand are designated UTM Zones 59 South and 60 South. Directional designations are here abbreviated, for example, 30N in place of 30 North.

Distances in the UTM system are specified in meters north and east of a zone origin (Figure 3-45). The y values are known as *UTM northings*, and increase in a northerly direction. The x values are referred to as *UTM eastings* and increase in an easterly direction.

The origins of the UTM coordinate system are defined differently depending on whether the zone is north or south of the Equator. In either case, the UTM coordinate system is defined so that all coordinates are positive within the zone. Zone easting coordinates are all greater than zero because the central meridian for each zone is assigned an easting value of 500,000 meters. This effectively places the origin (E = 0) at a point 500,000 meters west of the central meridian. All zones are less than 1,000,000 meters wide, ensuring that all eastings will be positive.

The Equator is used as the northing origin for all north zones. Thus, the Equator is assigned a northing value of zero for north zones. This avoids negative coordinates, because all of the UTM north zones are defined to be north of the Equator.

Universal Transverse Mercator zones south of the Equator are slightly different than those north of the Equator (Figure 3-46). South zones have a *false northing* value added to ensure all coordinates within a zone

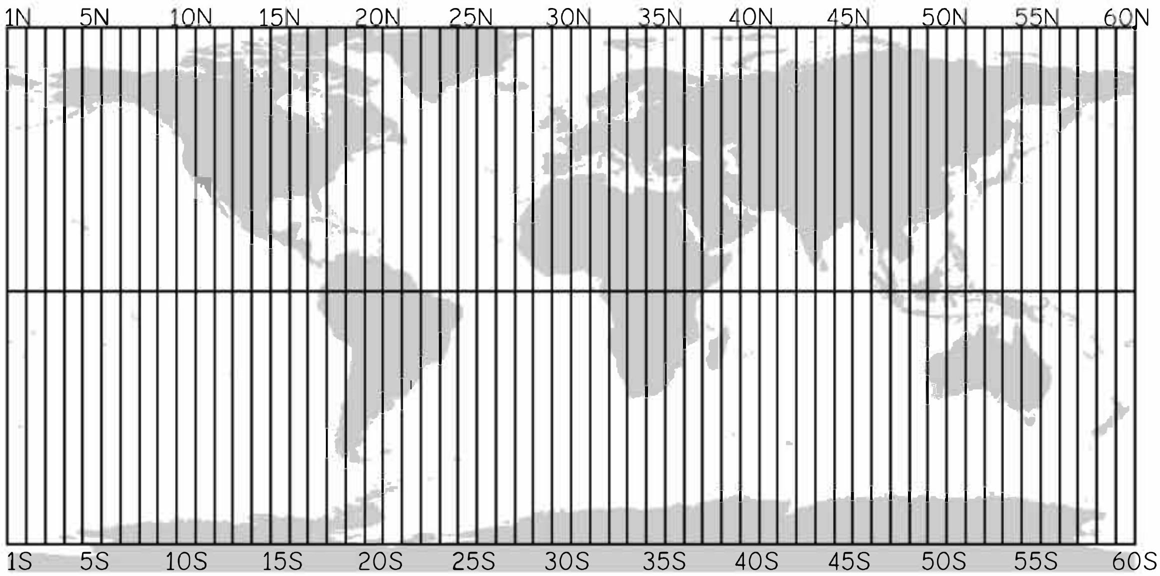


Figure 3-44: UTM zone boundaries and zone designators. Zones are six degrees wide and numbered from 1 to 60 from the International Date Line, 180°W. Zones are also identified by their position north and south of the Equator, e.g., Zone 7 North, Zone 16 South.

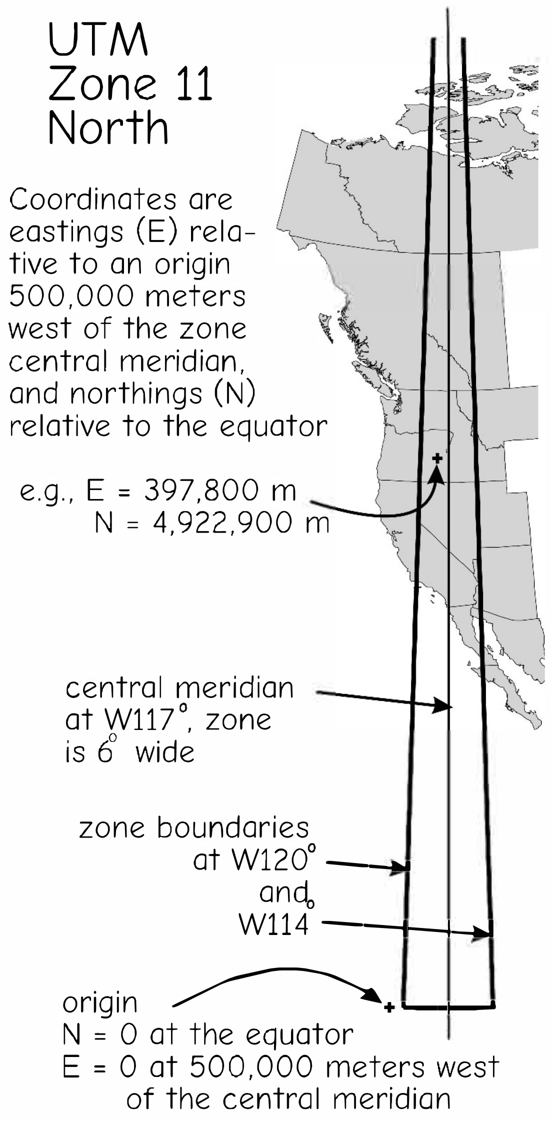


Figure 3-45: UTM zone 11N. The zone origin is on the Equator, with a false easting of 500,000 to ensure positive coordinates throughout the zone.

are positive. UTM coordinate values increase as one moves from south to north in a projection area. If the origin were placed at the Equator with a value of zero for south zone coordinate systems, then all the northing values would be negative. An offset is applied by assigning a false northing, a non-zero value, to an origin or other appropriate location. For UTM south zones, the northing values at the Equator are set to equal 10,000,000 meters. Because the distance from the Equator to the most southerly point in a UTM south zone is less than 10,000,000 meters, this assures that all northing coordinate values will be positive within each UTM south zone (Figure 3-46).

The UTM coordinate system is common for data and study areas spanning large regions, for example, several State Plane zones. Many data from U.S. federal government sources are in a UTM coordinate system because many agencies manage large areas. Many state government agencies in the United States distribute data in UTM coordinate systems because the entire state

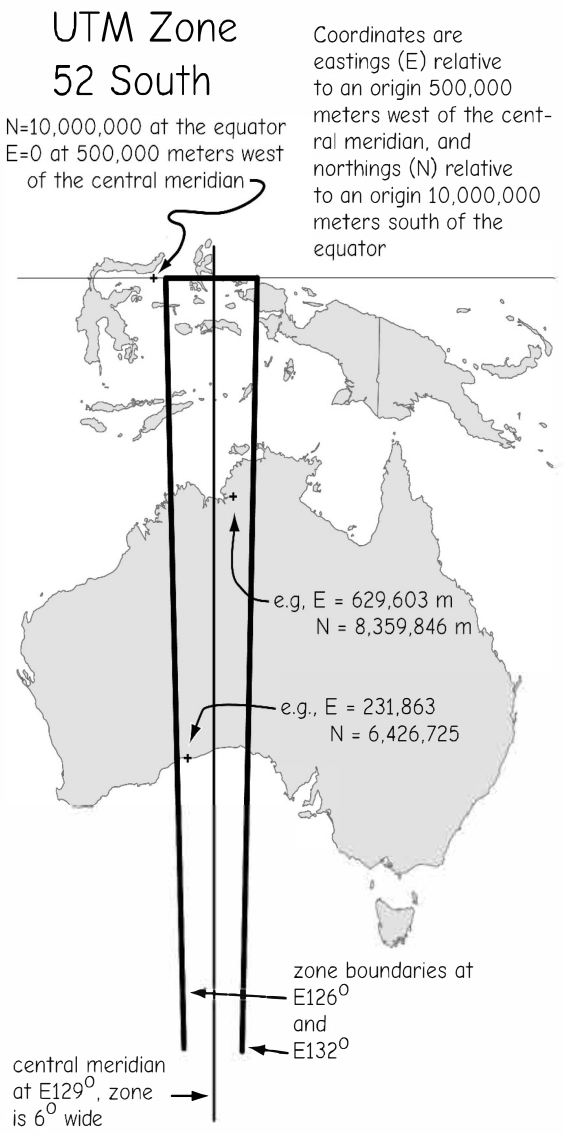


Figure 3-46: UTM south zones are defined to maintain positive northing and easting values within the zone. To that end, a false northing of 10,000,000 is applied to the Equator, and a false easting of 500,000 is applied to the central meridian.

fits predominantly or entirely into one UTM zone.

As noted before, all data for an analysis area must be in the same coordinate system if they are to be analyzed together. If not, the data will not co-occur as they should. The large width of the UTM zones accommodates many large-area analyses, and many states, national forests, or multicounty agencies have adopted the dominant UTM coordinate system as a standard. States that fall predominantly or entirely within a zone often adopt a UTM zone for much statewide data, e.g., Utah and UTM zone 12 (Figure 3-47).

We must note that the UTM coordinate system is not always compatible with regional analyses. Because coordinate values are discontinuous across UTM zone boundaries, analyses are difficult across these boundaries. UTM zone 15 is a different coordinate system than UTM zone 16. The state of Wisconsin approximately straddles these two zones, and the state of Georgia straddles zones 16 and 17. If a uniform, statewide coordinate system is required, the choice of zone is not clear, and either one or the other of these zones must be used, or some compromise projection must be chosen. For example, statewide analyses in Georgia and in Wisconsin are often conducted using UTM-like systems that involve moving the central meridian to near the center of each state.

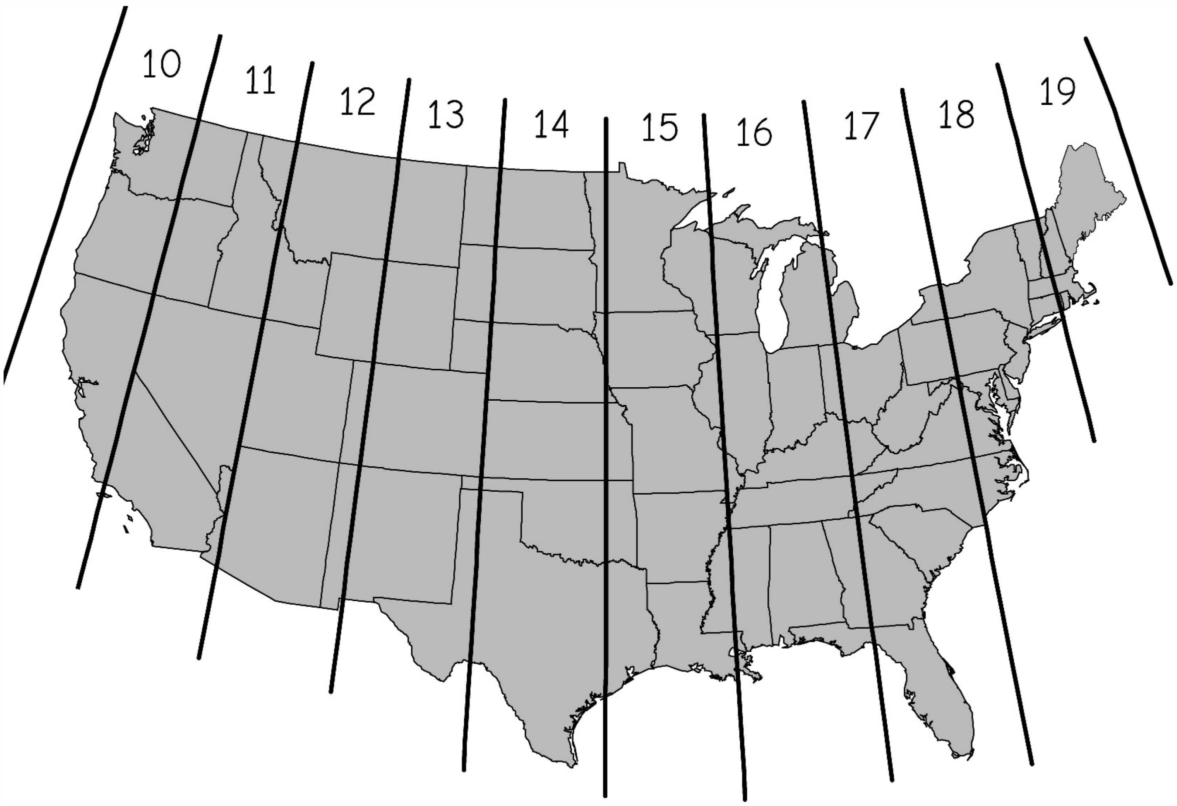


Figure 3-47: UTM zones for the lower 48 contiguous states of the United States of America. Each UTM zone is 6 degrees wide. All zones in the Northern Hemisphere are north zones, e.g., Zone 10 North.

National Coordinate Systems

Many governments have adopted a standard project for nationwide data, particularly small and midsized countries where distortion is limited across the spanned distances.

Many European countries have standard map projections covering a national extent; for example, Belgium, Estonia, and France each have different Lambert Conformal Conic projections defined for use on standard nation-spanning maps and data sets, while Germany, Bulgaria, Croatia, and Slovenia use a specialized modification of the transverse Mercator projection. Some countries adopt specific Universal Transverse Mercator projections, including Norway, Portugal, and Spain. Specifications of these projection parameters may be found in the respective national standard documents.

Larger countries may not have a specific or unified set of standard, nationwide projections, particularly for GIS data, because distortion is usually unavoidably large when spanning great distances across both latitudes and longitudes in the same map. There is simply no single projection that faithfully represents distances, areas, or angles across the entire country, so more constrained projections are used for analysis, and the results aggregated to larger areas.

Continental and Global Projections

There are map projections that are commonly used when depicting maps of the world. Directions, distances, and areas are typically not measured or computed on them, as distortions are too great. Most worldwide projections are used for visualization, but not quantitative analysis.

There are a number of projections that have been widely used for the world. These include variants of the Mercator, Goode, Mollweide, and Miller projections, among others. There is a trade-off that must be made in global projections, between a continuous map surface and distortion.

Distortion in world maps may be reduced by using a cut or interrupted surface. Different projection parameters or surfaces may be specified for different parts of the globe. Projections may be mathematically constrained to be continuous across the area mapped.

Figure 3-48 illustrates an interrupted projection in the form of a Goode homolosine. This projection is based on a sinusoidal projection and a Mollweide projection. These two projection types are merged at parallels of identical scale. The parallel of identical scale in this example is set near the midnorthern latitude of 44° 40′ N.

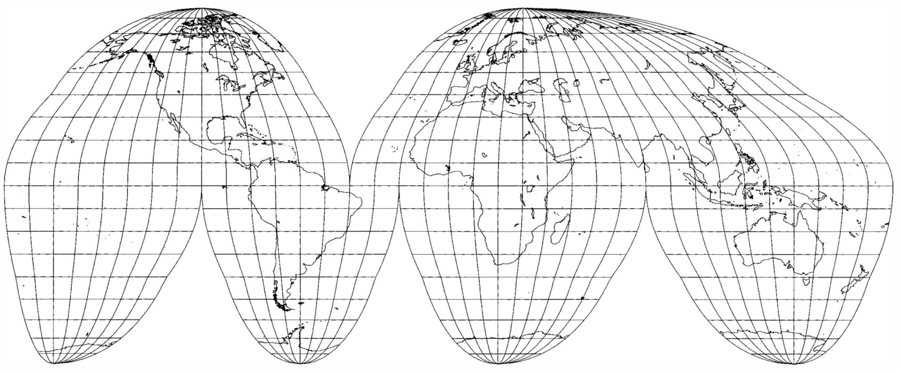


Figure 3-48: A Goode homolosine projection. This is an example of an interrupted projection, often used to reduce some forms of distortion when displaying the entire Earth surface (from Snyder and Voxland, 1989).

Conversion Among Coordinate Systems

Conversion from one projected coordinate system to another requires using the inverse and forward projection equations, described in an earlier section, passing through the geographic coordinate set. This allows a flexible conversion between any two projections, given our requirement that both the forward and inverse, or "backward" projection equations are specified for any map projection. For example, given a coordinate pair in the State Plane system, you may calculate the corresponding geographic coordinates. You may then apply a formula that converts geographic coordinates to UTM coordinates for a specific zone using another set of equations. Since the backward and forward projections from geographic to projected coordinate systems are known, we may convert among most coordinate systems by passing through a geographic system (Figure 3-49, a).

Care must be taken when converting among projections that use different datums. If appropriate, we must insert a datum transformation when converting from one projected coordinate system to another (Figure 3-49, b). A datum transformation, described earlier in this chapter, is a calculation of the change in geographic coordinates when moving from one datum to another.

Users of GIS software should be careful when applying coordinate projection tools

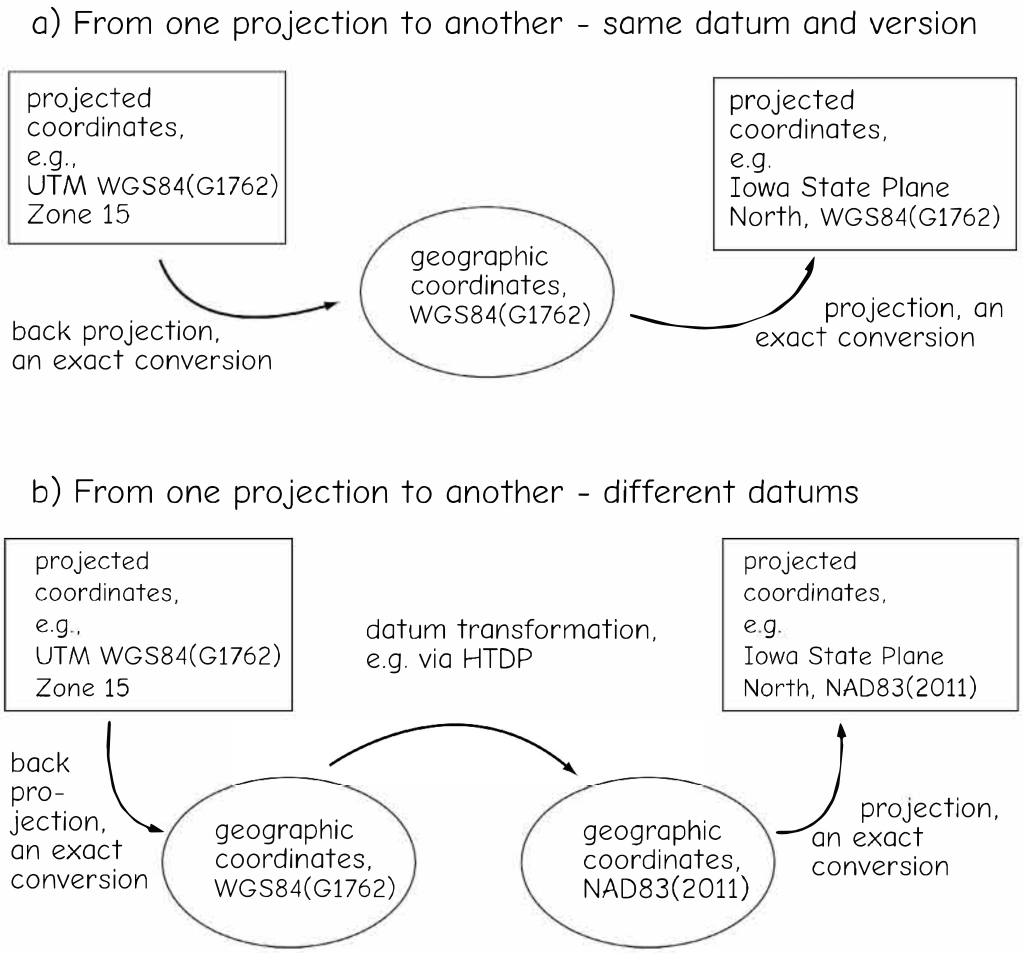


Figure 3-49: We may project between most coordinate systems via the back (or inverse) and forward projection equations. These calculate exact geographic coordinates from projected coordinates (a), and then new projected coordinates from the geographic coordinates. We must insert an extra step when a projection conversion includes a datum change. A datum transformation must be used to convert from one geodetic datum to another (b).

because the datum transformation may be omitted, or an inappropriate datum manually or automatically selected. For some software, the projection tool does not check or maintain information on the datum of the input spatial layer. This will often lead to an inappropriate or no datum transformation, and the output from the projection will be in error. Often these errors are small relative to other errors, for example, spatial imprecision in the collection of the line or point features. As shown in Figure 3-23, errors between NAD83(1986) and NAD83(CORS96) may be less than 10 cm (4 inches) in some regions, often much less than the average spatial error of the data themselves. However, errors due to ignoring the datum transformation may be quite large, for example, tens to hundreds of meters between NAD27 and most versions of NAD83, and errors of up to a meter are common between recent versions of WGS84/ITRF and NAD83. Given the sub-meter accuracy of many new GPS and other GNSS receivers used in data collection, datum transformation error of one meter is significant. As data collection accuracy improves, users develop applications based on those accuracies, so datum transformation errors should be avoided in all cases.

The Public Land Survey System

For the benefit of GIS practitioners in the United States, we must cover one final land designation system, known as the *Public Land Survey System*, or PLSS. The PLSS is not a coordinate system, but PLSS points are often used as reference points in the United States, so the PLSS should be well understood for work there.

The PLSS divided lands by north-south lines, 6 miles apart, running parallel to a principal meridian. East-west lines were surveyed perpendicular to these north-south lines, also at six mile intervals. These lines form square townships. Each township was further subdivided into 36 sections, each section approximately a mile on a side. Each section was subdivided further, to quarter-sections (one-half mile on a side), or sixteenth sections (one-quarter mile on a side). Sections were numbered in a zigzag pattern from one to 36, beginning in the northeast corner (Figure 3-50).

The PLSS is a standardized method for designating and describing the location of land parcels. It was used for the initial surveys over most of the United States after the early 1800s; therefore, nearly all land outside the original thirteen colonies uses the PLSS. An approximately uniform grid system was established across the landscape, with periodic adjustments incorporated to account for the anticipated error. Parcels were designated by their location within this grid system.

The PLSS was developed for a number of reasons. First, it was seen as a method to remedy many of the shortcomings of *metes and bounds* surveying, the most common method for surveying prior to the adoption of the PLSS. Metes and bounds describe a parcel relative to features on the landscape, sometimes supplemented with angle or distance measurements. Metes and bounds was used in colonial times, but parcel descriptions were often ambiguous. Subdivided parcels were often poorly described, and hence the source of much litigation, ill will, and many questionable real estate transactions.

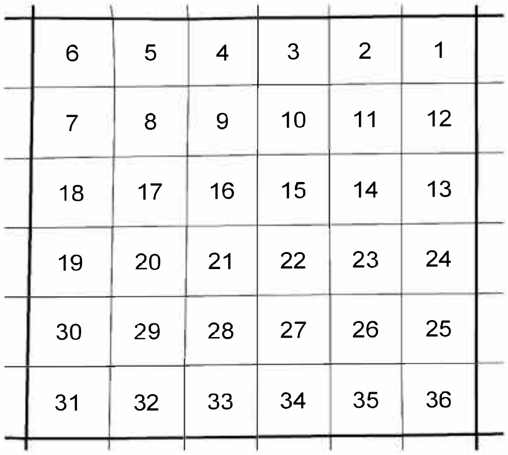


Figure 3-50: Typical layout and section numbering of a PLSS township.

The U.S. government needed a system that would provide unambiguous descriptions of parcels in unsettled territories west and south of the original colonies. The federal government saw public land sales as a way to generate revenue, to pay revolutionary war veterans, to expand the country, and to protect against encroachment by European powers. Parcels could not be sold until they were surveyed, so the PLSS was created. Land surveyed under the PLSS can be found in 30 states, including Alaska and most of the midwestern and western United States. Lands in the original 13 colonies, as well as West Virginia, Tennessee, Texas, and Kentucky were not surveyed under the PLSS system.

Surveyors typically marked the section corners and quarter-corners while running survey lines. Points were marked by a number of methods, including stone piles, pits, blaze marks chiseled in trees, and pipes or posts sunk in the ground.

Because the primary purpose of the PLSS survey was to identify parcels, lines and corner locations were considered static on completion of the survey, even if the corners were far from their intended location. Survey errors were inevitable given the large areas and number of different survey parties involved. Rather than invite endless dispute and readjustment, the PLSS specifies that boundaries established by the appointed PLSS surveyors are unchangeable, and that township and section corners must be accepted as true. The typical section contains approximately 640 acres, but due in part to errors in surveying, sections larger than 1200 acres and smaller than 20 acres were also established (Figure 3-51).

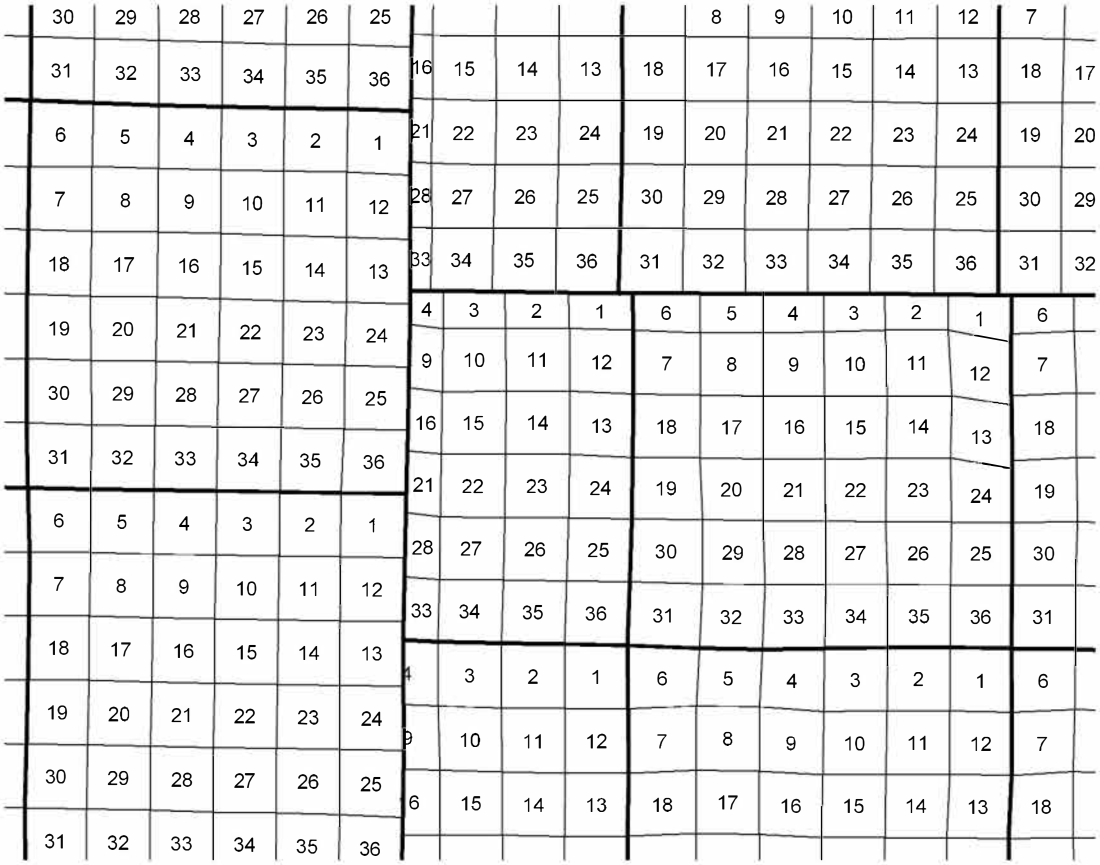


Figure 3-51: Example of variation in the size and shape of PLSS sections. Most sections are approximately one mile square with section lines parallel or perpendicular to the primary meridian, as illustrated by the township in the upper left of this figure. However, adjustments due to different primary meridians, different survey parties, and errors result in irregular section sizes and shapes.

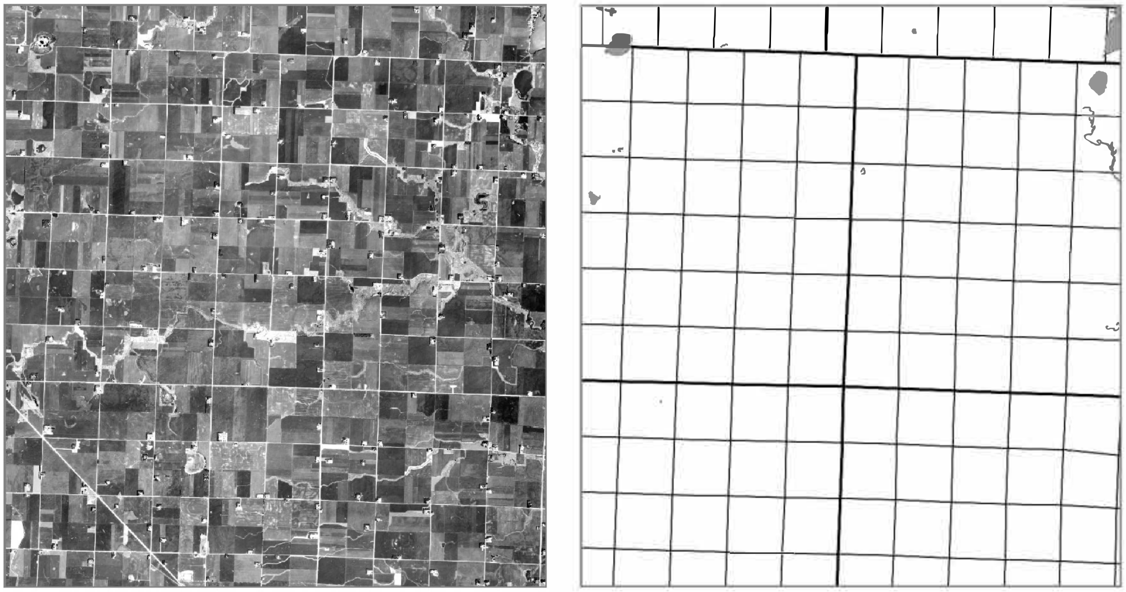


Figure 3-52: PLSS lines are often visible on the landscape. Roads (light lines on the image, above left) often follow the section and township lines (above right).

The PLSS is important today for several reasons. First, since PLSS lines are often property boundaries, they form natural corridors in which to place roads, powerlines, and other public services; they are often evident on the landscape (Figure 3-52). Many road intersections occur at PLSS corner points, and these can be viewed and referenced on many maps or imagery used for GIS database development efforts. Thus, the PLSS often forms a convenient system to coregister GIS data layers. PLSS corners and lines are often plotted on government maps (e.g., 1:24,000 quads) or available as digital data (e.g., National Cartographic Information Center Digital Line Graphs). Further, PLSS corners are sometimes resurveyed using high precision methods to provide property line control, particularly when a GIS is to be developed (Figure 3-53). These points may be useful to properly locate and orient spatial data layers on the Earth's surface.

Figure 3-53: A PLSS corner that has been surveyed and marked with a monument. This monument shows the physical location of a section corner. These points are often used as control points for further spatial data development.

Summary

In order to enter coordinates in a GIS, we need to uniquely define the location of all points on Earth. We must develop a reference frame for our coordinate system, and locate positions on this system. Since the Earth is a curved surface and we work with flat maps, we must somehow reconcile these two views of the world. We define positions on the globe via geodesy and surveying. We convert these locations to flat surfaces via map projections.

We begin by modeling the Earth's shape with an ellipsoid. An ellipsoid differs from the geoid, a gravitationally defined Earth surface, and these differences caused some early confusion in the adoption of standard global ellipsoids. There is a long history of ellipsoidal measurement, and we have arrived at our best estimates of global and regional ellipsoids after collecting large, painstakingly developed sets of precise surface and astronomical measurements. These measurements are combined into datums, and these datums are used to specify the coordinate locations of points on the surface of the Earth.

Map projections are a systematic rendering of points from the curved Earth surface onto a flat map surface. While there are many purely mathematical or purely empirical map projections, the most common map projections used in GIS are based on developable surfaces. Cones, cylinders, and planes are the most common developable surfaces. A map projection is constructed by passing rays from a projection center through both the Earth surface and the developable surface. Points on the Earth are projected along the rays and onto the developable surface. This surface is then mathematically unrolled to form a flat map.

Standard sets of projections are commonly used for spatial data in a GIS. In the United States, the UTM and State Plane coordinate systems define a standard set of map projections that are widely used. Other map projections are commonly used for continental or global maps, and for smaller maps in other regions of the world.

A datum transformation is often required when performing map projections. Datum transformations account for differences in geographic coordinates due to changes in the shape or origin of the spheroid, and in some cases to datum adjustments. Datum transformation should be applied as a step in the map projection process when input and output datums differ.

A system of land division known as the Public Land Survey System (PLSS) was established in the United States. This is not a coordinate system, but rather a method for unambiguously and systematically defining parcels of land based on regularly spaced survey lines in approximately north-south and east-west directions. Intersection coordinates have been precisely measured for many of these survey lines, and are often used as a reference grid for further surveys or land subdivision.

Suggested Reading

Bossler, J.D. (2002). Datums and geodetic systems. In J. Bossler (Ed.), *Manual of Geospatial Technology.* London: Taylor and Francis.

Brandenburger, A.J., Gosh, S K. (1985). The world's topographic and cadastral mapping operations. *Photogrammetric Engineering and Remote Sensing*, 51:437-444.

Burkholder, E.F. (1993). Computation of horizontal/level distances. *Journal of Surveying Engineering*, 117:104-119.

Colvocoresses, A.P. (1997). The gridded map. *Photogrammetric Engineering and Remote Sensing*, 63:371-376.

Doyle, F.J. (1997). Map conversion and the UTM Grid. *Photogrammetric Engineering and Remote Sensing*, 63:367-370.

Elithorpe, J.A.Jr., Findorff, D.D. (2009). *Geodesy for Geomatics and GIS Professionals*. Acton: Copley Custom Textbooks.

Featherstone, W.E., Kuhn, M. (2006). Height systems and vertical datums: a review in the Australian context. *Journal of Spatial Science*, 51:21-41.

Flacke, W., Kraus, B. (2005). *Working with Projections and Datum Transformations in ArcGIS: Theory and Practical Examples*. Norden: Points Verlag.

Habib, A. (2002). Coordinate transformation. In J. Bossler (Ed.), *Manual of Geospatial Technology*. London: Taylor and Francis.

Iliffe, J.C., Lott, R. (2008). *Datums and Map Projections for Remote Sensing, GIS, and Surveying*. 2nd ed. Boca Raton: CRC Press.

International Association of Oil and Gas Producers (2016). *Coordinate Conversion and Transformations including Formulas. Geomatics Guidance Note Number 7, Part 12*. www.epsg.org.

Janssen, V. (2009). Understanding coordinate reference systems, datums, and transformations. *International Journal of Geoinformatics*, 5:41-53.

Keay, J. (2000). *The Great Arc*. New York: Harper Collins.

Leick, A. (1993). Accuracy standards for modern three-dimensional geodetic networks. *Surveying and Land Information Systems,* 53:111-127.

Maling, D.H. (1992). *Coordinate Systems and Map Projections*. London: George Phillip.

Meyer, T.H., Roman, D.H., Zilkoski, D.B. (2006). What does height really mean? Part III: Height systems. *Surveying and Land Information Systems,* 66:149-160.

Milbert, D. (2008). An analysis of the NAD83(NSRS2007) National Readjustment. Downloaded 9/12/2011 from http://www.ngs.noaa.gov/PUBS_LIB/NSRS2007

National Geospatial-Intelligence Agency (NGA), TR8350.2 World Geodetic System 1984, Its Definition and Relationship with Local Geodetic Systems. http://earth-info.nga.mil/GandG/publications/tr8350.2/tr8350_2.html

NOAA Manual NOS NGS 5. State Plane Coordinate System of 1983. http://www.ngs.noaa.gov/PUBS_LIB/ManualNOSNGS5.pdf

Schuh, H., Behrend, B. (2012). VLBI: A fascinating technique for geodesy and astronomy. *Journal of Geodynamics*, 61:68-80.

Schwartz, C.R. (1989). *North American Datum of 1983, NOAA Professional Paper NOS 2*. Rockville: National Geodetic Survey.

Smith, D.S., Roman, D., Hilla, S. (2017). *NOAA Technical Report NOS NG62*. National Oceanic and Atmospheric Administration.

Smith, J. (1997). *Introduction to Geodesy: The History and Concepts of Modern Geodesy*. New York: Wiley.

Snay, R.A., Soler, T. (1999). Modern terrestrial reference systems, part 1. *Professional Surveyor*, 19:32-33.

Snay, R.A. Soler, T. (2000). Modern terrestrial reference systems, part 2. The evolution of NAD83. *Professional Surveyor*, 20:16-18.

Snay, R.A., Soler, T. (2000). Modern terrestrial reference systems, part 3. WGS84 and ITRS. *Professional Surveyor*, 20:24-28.

Snay, R.A., Soler, T. (2000). Modern terrestrial reference systems, part 4. Practical considerations for accurate positioning. *Professional Surveyor*, 20:32-34.

Snyder, J. (1993). *Flattening the Earth: Two Thousand Years of Map Projections*. Chicago: University of Chicago Press.

Snyder, J.P. (1987). *Map Projections, A Working Manual, USGS Professional Paper No. 1396*. Washington, D.C.: United States Government Printing Office.

Snyder, J.P., Voxland, P.M. (1989). *An Album of Map Projections, USGS Professional Paper No. 1453*. Washington D.C.: United States Government Printing Office.

Sobel, D. (1995). *Longitude*. New York: Penguin Books.

Soler, T., Snay, R.A.(2004). Transforming positions and velocities between the International Terrestrial Reference Frame of 2000 and the North American Datum of 1983. *Journal of Surveying Engineering*, 130:49-55.

Tobler, W.R. (1962). A classification of map projections. *Annals of the Association of American Geographers*, 52:167-175.

U.S. Coast and Geodetic Survey Special Publication 235. The State Coordinate Systems. http://www.ngs.noaa.gov/PUBS_LIB/publication235.pdf

Van Sickle, J. (2010). *Basic GIS Coordinates, 2nd Edition*. Boca Raton: CRC Press.

Vanicek, P., Steeves, R.H. (1996). Transformation of coordinates between two horizontal geodetic datums. *Journal of Geodesy*, 70:740-745.

Welch, R., Homsey, A. (1997). Datum shifts for UTM coordinates. *Photogrammetric Engineering and Remote Sensing*, 63:371-376.

Wolf, P.R., Ghilani, C.D. (2002). *Elementary Surveying.* 10th ed. Upper Saddle River:Prentice-Hall.

Yang, Q., Snyder, J.P., Tobler, W.R. (2000). *Map Projection Transformation: Principles and Applications.* London: Taylor & Francis.

Zenk, D. (2014). Correct use of NAD83 realizations and geoid model. *Minnesota Surveyor,* 22:16-18.

Zilkoski, D., Richards, J., Young, G. (1992). Results of the general adjustment of the North American Vertical Datum of 1988. *Surveying and Land Information Systems*, 53:133-149.

Study Questions

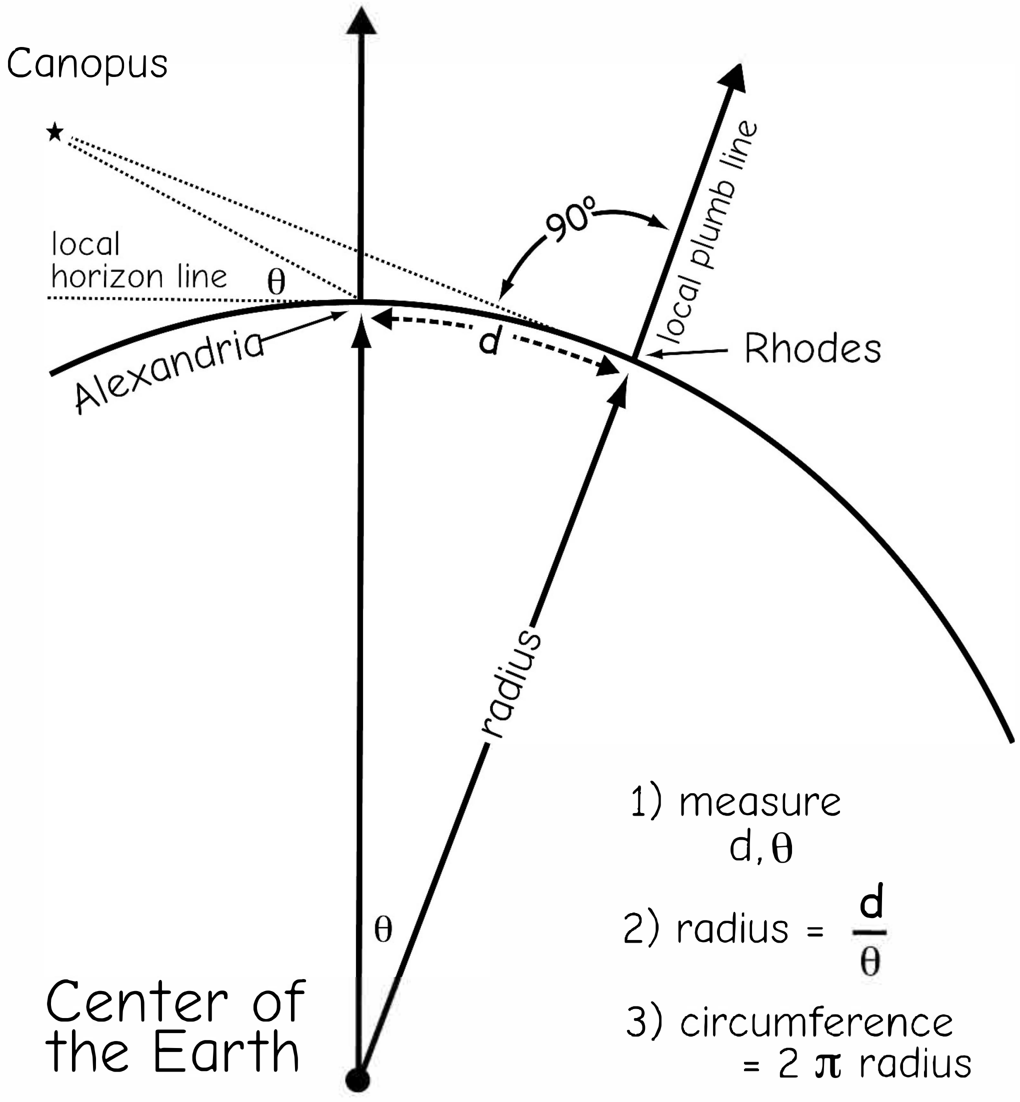
3.1 - Describe how Eratosthenes estimated the circumference of the Earth. What value did he obtain?

3.2 - Assume the Earth is approximately a sphere (not an ellipsoid). Also assume you've repeated the measurements of Poseidonius, shown in the figure below. What is your estimate of the radius of the Earth's sphere given the following distance/angle pairs. Note that the distances are given below in meters, and angle in degrees, and calculators or spreadsheets may require you enter angles in radians for trigonometric functions (1 radian = 57.2957795 degrees):

a) angle $\theta = 1^{o}\ 18'\ 45.79558''$, distance = 146,000 meters

b) angle $\theta = 0^{o}\ 43'\ 32.17917''$, distance = 80,500 meters

c) angle $\theta = 0^{o}\ 3'\ 15.06032''$, distance = 6,000 meters



3.3 - Assume the Earth is approximately a sphere (not an ellipsoid). Also assume you've repeated the measurements of Poseidonius. What is your estimate of the radius of the Earth's sphere given the following distance/angle pairs. Note that the distances are given in meters, and angle in degrees, and calculators or spreadsheets may require you enter angles in radians for trigonometric functions (1 radian = 57.2957795 degrees):

a) angle = 2° 59' 31.33325", distance = 332,000 meters

b) angle = 9° 12' 12.77201", distance = 1,020,708 meters

c) angle = 1° 2' 12.15566", distance = 115,200 meters

3.4 - What is an ellipsoid? How does an ellipse differ from a sphere? What is the equation for the flattening factor?

3.5 - Provide three reasons why there have been various estimates for Earth's ellipsoid radii.

3.6 - Define the geoid. Tell how it differs from the ellipsoid, and from the surface of the Earth. Describe how we measure the position of the geoid.

3.7 - Define a parallel or meridian in a geographic coordinate system. Describe where the zero lines occur.

3.8 - How does magnetic north differ from the geographic North Pole?

3.9 - Define a datum. Describe how datums are developed.

3.10 - Why are there multiple datums, even for the same place on Earth? Define what we mean when we say there is a datum shift.

3.11 - What is a triangulation survey, and what is a bench mark?

3.12 - Why do we not measure vertical heights relative to mean sea level?

3.13 - What is the difference between an orthometric height and a dynamic height?

3.14 - Use the NCAT software available from the U.S. NOAA/NGS website (https://www.ngs.noaa.gov/NCAT/) to fill the following table. Note that all of these points are in the continental United States (CONUS) and longitudes are west, but entered as positive numbers.

		NAD27		NAD83(86)		HPGN	
Pnt	State	latitude	longitude	latitude	longitude	latitude	longitude
1	Calif. (S)	32°44'15"	117°09'42"	32°44'15.1827"	117°09'45.1202"	32°44'15.1870"	117°09'45.1201"
3	Wisconsin	43°07'59"	89°20'11"	43°07'58.9806"	89°20'11.4226"		
5	Colorado	40°00'00"	105°16'01"			40°00'00.0068"	105°16'02.9711"
7	Wash. D.C.			38°51'10.4052"	77°02'19.9165"	38°51'10.4064"	77°02'19.9041"

3.15 - Use the NCAT software available from the U.S. NOAA/NGS website (https://www.ngs.noaa.gov/NCAT/) to fill the following table. Note that all of these points are in the continental United States (CONUS), and longitudes are west, but entered as positive numbers.

		NAD27		NAD83(86)		HPGN	
Pnt	State	latitude	longitude	latitude	longitude	latitude	longitude
2	Washington	47°27'55"	122°18'06"	47°27'54.3574"	122°18'10.4453"	47°27'54.3642"	122°18'10.4366"
4	Texas	29°58'07"	95°21'31"	29°58'07.7975"	95°21'31.7705"		
6	Florida (S)	24°33'30"	81°45'19"			24°33'31.5216"	81°45'18.3362"
8	Maine			46•52'0.1524"	68°00'59.0974"	46°52'0.1580"	68°00'59.0995"

3.16 - Use the World Wide Web version or download and start the HTDP software from the U.S. NOAA/NGS site (at the time of this writing, http://www.ngs.noaa.gov/TOOLS/Htdp/Htdp.shtml), and complete the following table. Use the tool for a horizontal displacement between two dates. Enter epoch start and stop dates of January 1, 1986 and January 1, 2015, respectively. Specify a zero height or z for your datum transformation. Use the spherical Earth approximation formulas described in Chapter 2 when calculating the surface shift distance, in centimeters (cm), assuming a radius of 6,371 kilometers. Report the ground shift from 1986 to the 2015 time period.

	NAD83(2011) - 1986		NAD83(2011) - 2015		Surface shift distance (cm)	
Pnt	latitude	longitude	latitude	longitude	latitude	longitude
1	32°44'15"	117°09'42"	32°44'15.0292"	117°09'42.0298"	-90.2	-92.0
3	43°07'59"	89°20'11"	43°07'58.9954"	89°20'10.9973"		
5	40°00'00"	105°16'01"				
7			38°51'0.992"	77°02'20.9978"		

3.17 - Use the World Wide Web version or download and start the HTDP software from the U.S. NOAA/NGS site (at the time of this writing, http://www.ngs.noaa.gov/TOOLS/Htdp/Htdp.shtml), and complete the following table. Use the tool for a horizontal displacement between two dates. Enter epoch start and stop dates of January 1, 1986 and January 1, 2015, respectively. Specify a zero height or z for your datum transformation. Use the spherical Earth approximation formulas described in Chapter 2 when calculating the surface shift distance, in centimeters (cm), assuming a radius of 6,371 kilometers. Report the ground distance between points from the 1986 to the 2015 time period.

	NAD83(2011) - 1986		NAD83(2011) - 2015		Surface shift distance (cm)	
Pnt	latitude	longitude	latitude	longitude	latitude	longitude
2	47°27'55"	122°18'06"	47°27'55.0052"	122°18'05.9914"	-16.1	26.6
4	29°58'07"	95°21'31"	29°58'07.0002"	95°21'30.9981"		
6	24°33'30"	81°45'19"				
8			46°51'59.9983"	68°00'59.9970"		

3.18 - Use the VDatum software (available at the time of this writing at https://vdatum.noaa.gov/) to complete the table. Note that all longitudes are west, entered as negative numbers.

		NAD27			NAD83(2011)		
Pnt	State	latitude	longitude	elevation (m)	latitude	longitude	elevation (m)
1	Calif.	32°40'00"	-117°00'00"	200	32°40'00.1890"	-117°00'03.0990"	200.654
3	Washington	48°30'00"	-122°00'00"	200			
5	Maine	47°00'00"	-69°00'00"	200			
7	Florida	25°00'00"	-81°30'00"	1			

3.19 - Use the VDatum software (available at the time of this writing at https://vdatum.noaa.gov/) to complete the table. Note that longitudes are west, entered as negative numbers.

		NAD27			NAD83(2011)		
Pnt	State	latitude	longitude	elevation (m)	latitude	longitude	elevation (m)
2	Minnesota	45°00'00"	-95°00'00"	200	45°59'59.8684"	-95°00'01.0170"	200.169
4	Texas	30°15'00"	-97°45'00"	200			
6	Colorado	38°00'00"	-107°45'00"	3000			
8	N. Carolina	35°40'00"	-82°30'00"	500			

3.20 - Use the VDatum software to calculate the orthometric height change, in centimeters, for the listed NAD83(2011) geographic coordinates, when switching from the NAVD(geoid12A) as the source, to geoid 2009, 1999, and 1996 geoids respectively.,

Pnt	State	latitude	longitude	geoid12A elevation (m)	Δheight (cm), to geoid09	Δheight (cm), to geoid99	Δheight (cm), to geoid96
1	Calif.	32°40'00"	-117°00'00"	200	0	5.2	6.6
3	Washington	48°30'00"	-122°00'00"	200			
5	Maine	47°00'00"	-69°00'00"	200			
7	Florida	25°00'00"	-81°30'00"	1			

3.21 - Use the VDatum software to calculate the orthometric height change, in centimeters, for the listed NAD83(2011) geographic coordinates, when switching from the NAVD(geoid12A) as the source, to geoid 2009, 1999, and 1996 geoids respectively.

Pnt	State	latitude	longitude	geoid12A elevation (m)	Δheight (cm), to geoid09	Δheight (cm), to geoid99	Δheight (cm), to geoid96
2	Minnesota	45°00'00"	-95°00'00"	200	2.9	7.7	8.1
4	Texas	30°15'00"	-97°45'00"	200			
6	Colorado	38°00'00"	-107°45'00"	3000			
8	N. Carolina	35°40'00"	-82°30'00"	500			

3.22 - a) You wish to site a seaside hospital in San Diego, CA. Using gauge 9410170, report the NAVD88 elevation you wish to use as a threshold if you want the site to be at least 30 feet above the mean high water mark (the search function in the upper left corner of the tides website, described in the tides section of this chapter, will help speed the search for the station information. Then look for a tides and water levels, datums section).

b) What is the height difference between the gauge NAVD88 height and the Mean low low water mark for the station?

3.23 - a) You wish to dredge a channel near Naples, FL, near the NOAA tidal gauge station 8725110. You wish to maintain a channel depth of 30 feet below the mean sea level. What is the channel depth elevation expressed as an NAVD88 height, in feet? (the search function in the upper left corner of the tides website, described in the tides section of this chapter, will help find the station. Then look for a tides and water levels, datums section)

b) What is the mean high-high water mark, expressed in feet, as a NAVD88 height?

3.24 - What is a developable surface? What are the most common shapes for a developable surface?

3.25 - Look up the NGS control sheets for the following points, and record their horizontal and vertical datums, latitudes, and longitudes:

DOG, Maine, PID= PD0617

Key West GSL, Florida, PID=AA1645

Neah A, Washington, PID=AF8882

3.26 - Look up the NGS control sheets for the following points, and record their horizontal and vertical datums, latitudes, and longitudes:

Denver, Colorado, PID= KK1544

Loma East, CA, PID=AC6092

Austin CE, Texas, PID=DN7664

3.27 - Using the spheroid formula given in this chapter, calculate the great circle distance to the nearest kilometer for the control points in question 3.25 - above from:

- DOG to Neah A
- Key West to DOG
- Neah A to Key West

3.28 - Calculate the great circle distance for the control points in question 3.26 - above from:

- Denver to Loma East
- Denver to Austin CE
- Austin CE to Loma East

3.29 - Describe the State Plane coordinate system. What type of projections are used in a State Plane coordinate system?

3.30 - Define and describe the Universal Transverse Mercator coordinate system. What type of developable surface is used with a UTM projection? What are UTM zones, where is the origin of a zone, and how are negative coordinates avoided?

3.31 - What is a datum transformation? How does it differ from a map projection?

3.32 - Specify which type of map projection you would choose for each country, assuming you could use only one map projection for the entire country, the projection lines of intersection would be optimally placed, and you wanted to minimize overall spatial distance distortion for the country. Choose from a transverse Mercator, a Lambert conformal conic, or an Azimuthal:

Benin	Bhutan
Slovenia	Israel

3.33 - Specify which type of map projection you would choose for each country, assuming you could use only one map projection for the entire country, the projection lines of intersection would be optimally placed, and you wanted to minimize overall spatial distance distortion for the country. Choose from a transverse Mercator, a Lambert conformal conic, or an Azimuthal:

Chile	Nepal
Kyrgyzstan	The Gambia

3.34 - Describe the Public Land Survey System. Is it a coordinate system? What is its main purpose?

4 Maps, Data Entry, Editing, and Output

Building a GIS Database

Introduction

Spatial data entry and editing are frequent activities for many GIS users. A large number of coordinates is needed to represent features in a GIS, and each coordinate value must be entered into the GIS database. This is often painstakingly slow, even with automated techniques, and spatial data entry and editing take significant time for most organizations.

Most spatial data sources may be categorized as either *hardcopy* or *digital*. Hardcopy forms are any drawn, written, or printed documents, including hand-drawn maps, manually measured survey data, legal records, and coordinate lists with associated tabular data. Most historical spatial data were recorded on maps (Figure 4-1), and although not all maps are suitable for conversion to digital formats, many maps are. Much data were created from hardcopy sources in the early years of GIS via *digitizing*, the process of collecting

Figure 4-1: Maps have served to store geographic knowledge for at least the past 4,000 years. This early map of northern Europe shows approximate shapes and relative locations.

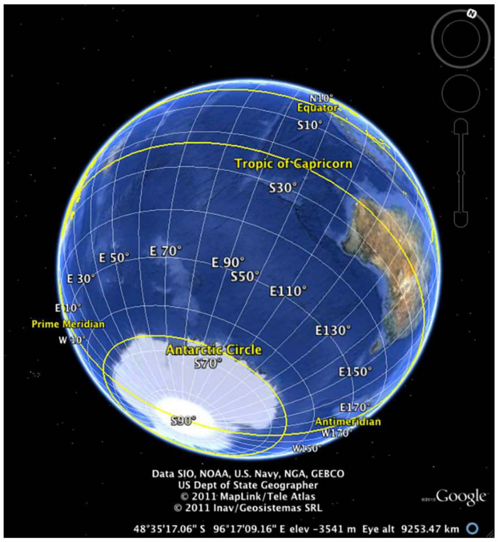


Figure 4-2: An example of commonly produced digital maps (courtesy Google).

digital coordinates. Digitizing is a common data entry method today, although primarily from satellite and aerial images.

Digital maps are an electronic, graphic depiction of spatial data, and are by far the most common map form today (Figure 4-2). Millions of electronic maps are generated each hour, composed on demand in response to web queries, on automobile navigations systems, and for commerce and advertising. These maps are flexible, easily customized, inexpensively distributed, and often dynamic.

Most maps, whether digital or hardcopy, contain several components (Figure 4-3). A *data area* or *pane* occupies the largest part of the map, and contains most of the depicted spatial data. A *neatline* is often included to provide a frame around all map elements, and *insets* may contain additional map elements. *Scalebars, legends*, titles, and other graphic elements such as a *north arrow* are often included. All maps have a *map scale*, defined as the ratio of the distance on the map to corresponding distance on the ground.

Maps often depict coordinate lines (Figure 4-4). When the lines represent constant latitude and longitude, a set of coordinate lines is called a *graticule* (Figure 4-4a). These lines may appear curved, depending on the map scale, the map coordinate system, and the location of the area on Earth's surface. Maps may also depict a *grid* con-

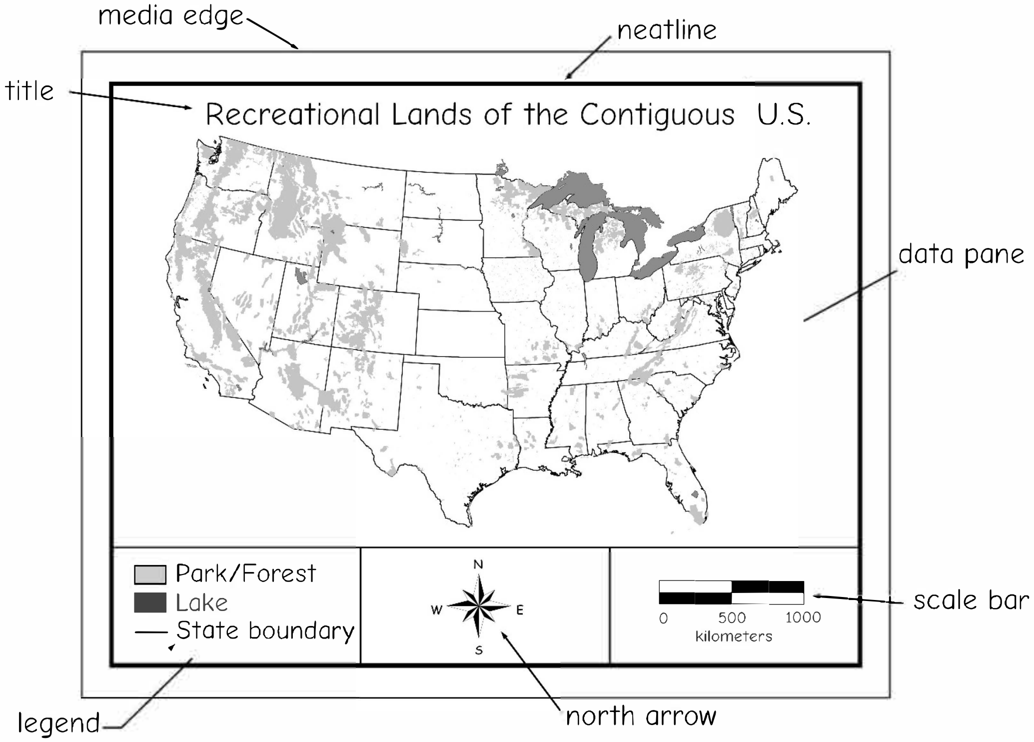


Figure 4-3: An example of a map and its components.

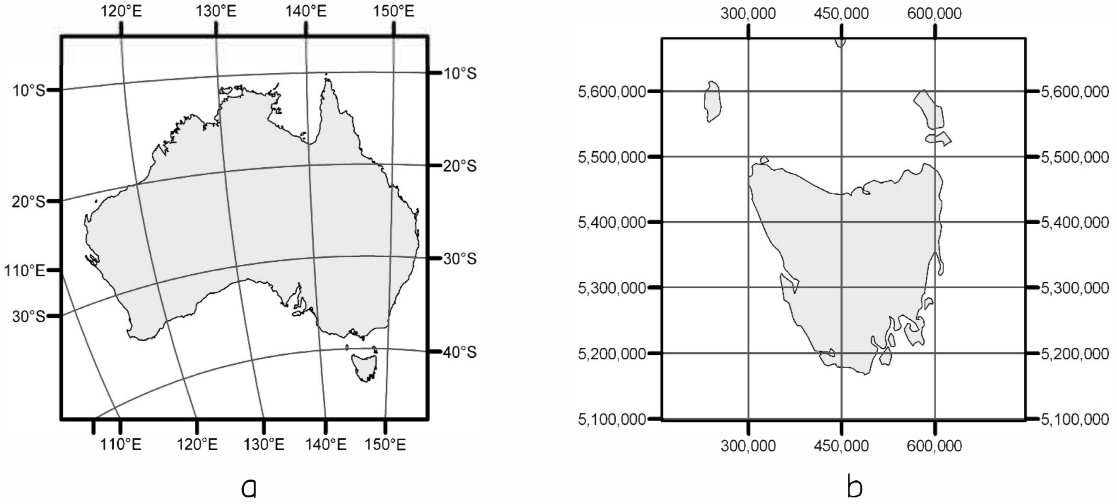


Figure 4-4: Maps often depict lines representing (a) a *graticule* of constant latitude and longitude or (b) a *grid* of constant X and Y coordinates.

sisting of lines of constant coordinates. Grid lines are typically drawn in both the x and y directions, and appear straight on most maps (Figure 4-4b). Graticules and grids are useful because they provide a reference against which location may be quickly estimated. Graticules are particularly useful for depicting the distortion inherent in a map projection, because they show how geographic north or east lines are deformed, and how this distortion varies across the map. Grids may establish a map-projected north, in contrast to geographic north, and may be useful when trying to navigate or locate a position on the map.

Historical and current images are valuable sources of geographic data, and although they are not maps, the line is becoming blurred, as aerial and satellite photographs become common backdrops for digital maps. Photographs do not typically provide an orthographic (flat, undistorted) view unless they are geometrically corrected, and houses, rivers, or features of interest are not explicitly identified. However, images are a rich source of geographic information, and standard techniques may be used to extract features through manual digitizing (described later in this chapter), or through image classification (described in Chapter 6).

Digital spatial data are those provided in a computer-compatible format. These include complete raster and vector data layers, text files, lists of coordinates, and digital images. Files and export formats can be used to transfer them to a local GIS system. Global Navigation Satellite Systems (GNSS), such as the U.S. Global Positioning System (GPS) and coordinate survey devices described in Chapter 5, are direct measurement systems that can be used to record coordinates in the field and report them directly into digital formats. Finally, many digital image sources are available, such as satellite or airborne images that are collected in a digital raster format, or hardcopy aerial photographs that have been scanned to produce digital images.

Hardcopy maps were an important source of geographic information, but now are primarily used for display. Most geographic information produced before 1980 was recorded in hardcopy form. Advances in optics, metallurgy, and industry during the 18th and 19th centuries allowed the mass production of precise surveying devices, allowing much of the world to be plotted on *cartometric* quality maps. These maps faithfully represent the relative position of objects.

Data entry from digital sources now dominates. Coordinates are increasingly captured via interpretation of digital image sources (these sources are described in Chapter 6) or collected directly in the field by satellite-based positioning services (Chapter 5).

Our objective in this chapter is to introduce spatial data entry via digitizing and coordinate surveying. We will also cover basic editing methods and data documentation, and rudimentary cartography and output.

Map Types

Many types of digital and hardcopy maps are produced, and the types are often referred to by the way features are depicted on the map. *Feature maps* are commonly used to map points, lines, or areas and provide nominal information (Figure 4-5, upper left). No attempt is made for symbols to represent true scale. A road may be plotted with a symbol defining the type of road, but the width of the road as plotted is not true to scaled size on the ground.

Choropleth maps depict quantitative information for areas. A mapped variable such as population density may be represented in the map (Figure 4-5, top right). Polygons define area boundaries, such as counties or census tracts. Each polygon is given a color, shading, or pattern corresponding to values for a mapped variable.

Dot density maps show quantitative data (Figure 4-5, bottom left). Dots or other point symbols are plotted to represent values. Dots are placed in the polygon such that the number of dots equals the total value for the

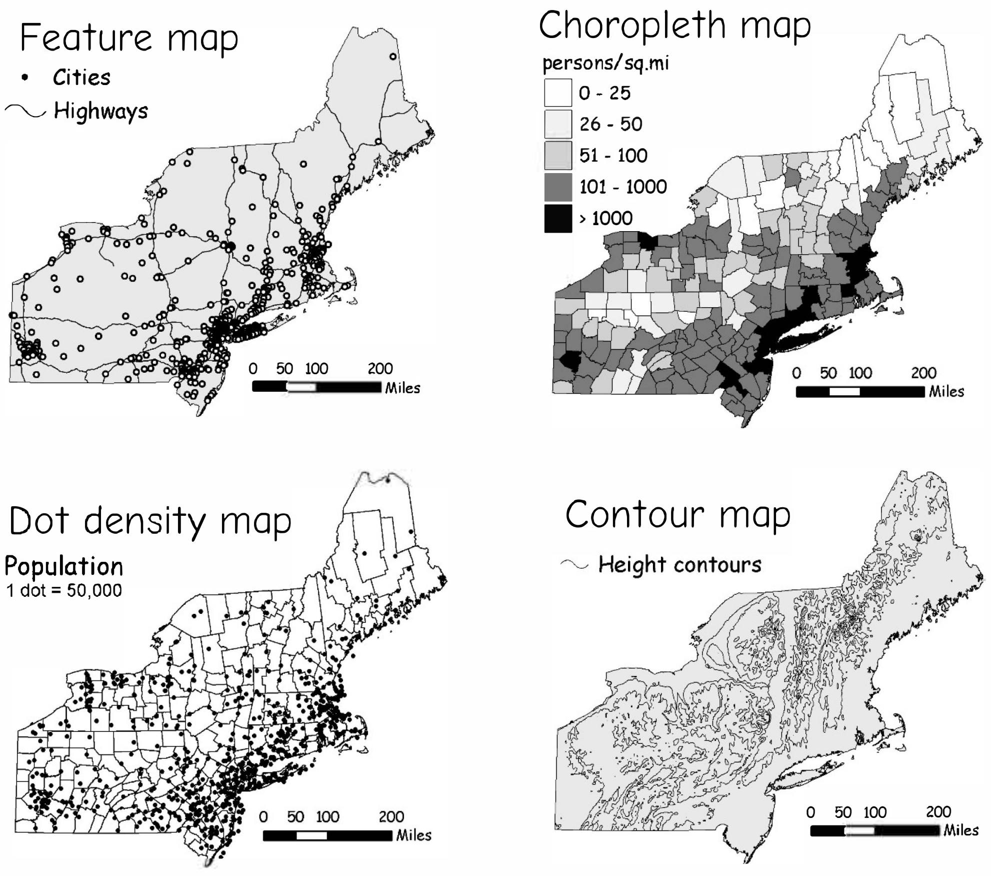


Figure 4-5:
Common hardcopy map types depicting the northeastern United States.

polygon. Note that the dots are typically placed randomly within the polygon area. Each dot on the example map represents 50,000 people; however, each point is not a city or other concentration of inhabitants. Note the position of points in the dot-density map relative to the city locations in the feature map directly above it in Figure 4-5.

Isopleth maps, also known as *contour maps*, display lines of equal value (Figure 4-5, bottom right). Isopleth maps are used to represent continuous surfaces. Rainfall, elevation, and temperature are features that are commonly represented using isopleth maps. A line on the isopleth map represents a specified value, for example, a 10°C isopleth defines the position on the landscape at that temperature. Lines typically do not cross, in that there cannot be two different temperatures at the same location. However, isopleths often depict elevation, and cliffs or overhanging terrain do have multiple elevations at the same location. In this case the lower elevations typically pass “under” the higher elevations, and the isopleth is labeled with the tallest height (Figure 4-6). Note that the isopleths are typically estimated surfaces, with the lines drawn based on measurements at a set of point locations; various methods for estimating isopleth lines from points are described in chapter 12.

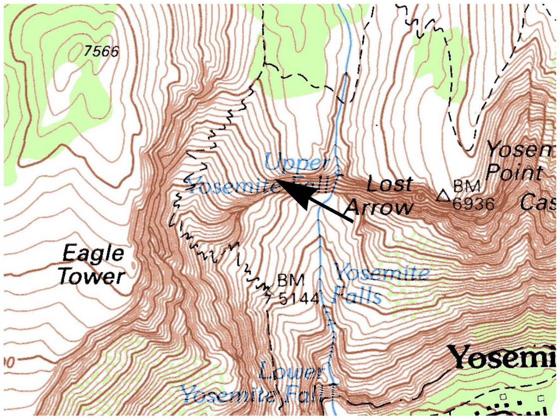


Figure 4-6: Lines on isopleth maps typically do not cross. However, as shown at the arrow in this image, lines may coincide when there is a common value. Here cliffs or overhangs result in converging isopleth lines.

Scale

All maps or digitally displayed data have a scale, a relationship between a distance on the map or screen and a corresponding distance projected on Earth. Map or screen scale is often reported as a distance conversion, such as one inch to a mile, meaning one inch on the map equals one mile on Earth. They may also be expressed as a unitless ratio, such as 1:24,000, indicating a unit distance on the display is equal to 24,000 units on Earth’s surface. Digital maps most often use a third method to report scale, as a bar or line of known distance, labeled on the map (Figure 4-7).

Note that depicting map scale was unambiguous when only hardcopy maps were produced. A written ratio or conversion, for example, 1 inch to the mile, was true because the features were fixed on paper or other physical media. When displaying data on-screen with a known dimension, the scale may be re-computed at any display size, because the computer can measure the displayed and data distances, and update the displayed scale in real time. However, displaying a fixed scale may be erroneous on an electronic document such as a pdf without georeferencing, because the scale is altered by zooming. This changes the magnification on an electronic display, without the ability to automatically recompute scale. One inch as displayed may not correspond to a mile. Fixed-scale digital documents should most often include a graphic scale bar, depicting an equivalent surface distance, for example, 1 km, embedded in the map.

The notion of large vs. small scale is often confused because scale implies a ratio or fraction. A larger ratio signifies a large-scale map, so a 1:24,000 scale map is considered large-scale relative to a 1:100,000 scale map. Many people mistakenly refer to a 1:100,000 scale map as larger scale than a 1:24,000 scale map because it covers a larger area. A 1:100,000 scale map that is 50 cm (20 in) on a side covers more ground than a 1:24,000 scale map that is 50 cm on a side. However, it is the size of the ratio or frac-

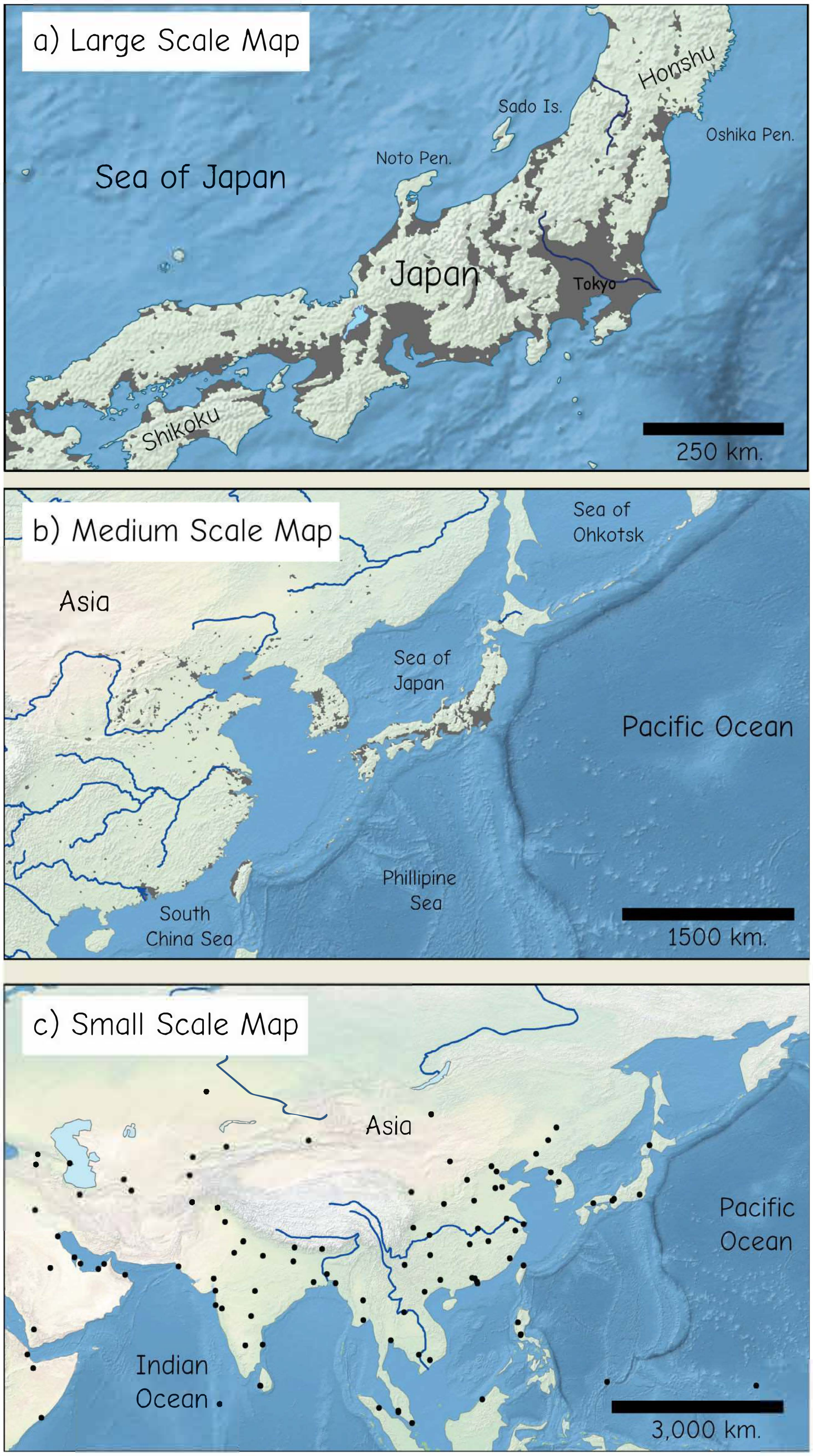


Figure 4-7: Coverage, relative distance, and detail change from larger scale (top) to smaller scale (bottom) maps.

tion, and not the area covered, that determines the map scale. It is helpful to remember that features appear larger on a large-scale map (Figure 4-7). It is also helpful to remember that large-scale maps of a given physical dimension often show more detail, but less area. Notice in Figure 4-7, the larger-scale map at the top shows details of Tokyo city. Tokyo shrinks in the successively smaller-scale maps, but large additional areas are covered. The larger the ratio (and smaller the denominator), the larger the map scale.

Because maps often report an average scale, and because there are upper limits on the accuracy with which data can be plotted on a map, large-scale maps generally have less geometric error than small-scale maps if the same methods were used to produce them. Small errors in measurement, plotting, printing, and paper deformation are magnified by the scale factor. These errors, which occur during map production, are magnified more on a small-scale map than a large-scale map.

Map and Data Generalization

Maps are abstractions of reality, as are spatial data in a GIS database. This abstraction introduces *map generalization*, the unavoidable approximation of real features when they are represented on a map. Not all the geometric or attribute detail of the physical world are recorded; only the most important characteristics are included. The set of features that are most important is subjectively defined and will differ among users. The mapmaker determines the set of features to place on the map, and selects the methods to collect and represent the shape and location of these features on the map.

The choice of data sources and digitizing methods will unavoidably set limits on the size and shape of features that may be represented. Consider mapping lakes, based on image data with a 250-meter cell size (Figure 4-8). The abstraction of the shoreline will not represent bays and peninsulas that are smaller than approximately 250 meters across, by conscious choice of the mapmakers. Small features will be missed, edge detail will be lost, and distances along boundaries will depend on the resolution of the source image.

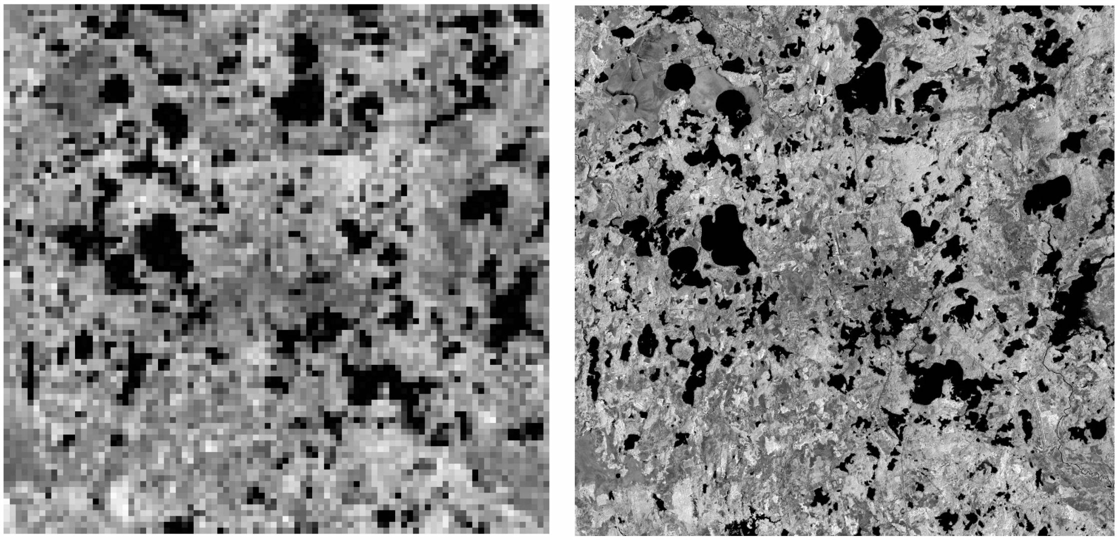


Figure 4-8: A mapmaker chooses the materials and methods used to produce a map, and so imposes a limit on spatial detail. Here, the choice of an input image with a 250 meter resolution (left) renders it impossible to represent all the details of the real lake boundaries (right). In this example, features smaller than approximately 250 meters on a side may not be faithfully represented on the map.

A finer resolution source, such as a 30-meter resolution, may more faithfully depict map detail, but may not be an appropriate choice. The finer resolution may be more expensive, difficult to reproduce, unavailable for the entire mapping area, or inappropriate because it does not show important features, for example, vegetation types or recent developments. Cartographers often must balance several factors in map design, and their choices inevitably lead to some form of map generalization.

Feature generalization is one common form of generalization. Feature generalization is a modification of features when representing them on a map. The geographic aspects of features are generalized because there are limits on the time, methods, or materials available when collecting geographic data. These limits also apply when compiling, displaying, or printing a map. These feature generalizations, depicted in Figure 4-9, may be classed as:

Fused: multiple features may be grouped to form a larger feature,

Simplified: boundary or shape details are lost or "rounded off",

Displaced: features may be offset to prevent overlap or to provide a standard distance between mapping symbols,

Omitted: Small features in a group may be excluded from the map, or

Exaggerated: standard symbol sizes are often chosen, for example, standard road symbol widths, which are much larger when scaled than the true road width.

Generalization is present at some level in almost every data layer or map, and should be recognized and evaluated for each data source in a GIS, including maps (Figure 4-10). If generalization results in omission or degradation of data beyond acceptable levels, then the analyst should switch to a larger-scale map if appropriate and available, or return to the field or original source materials to collect data at the required precision.

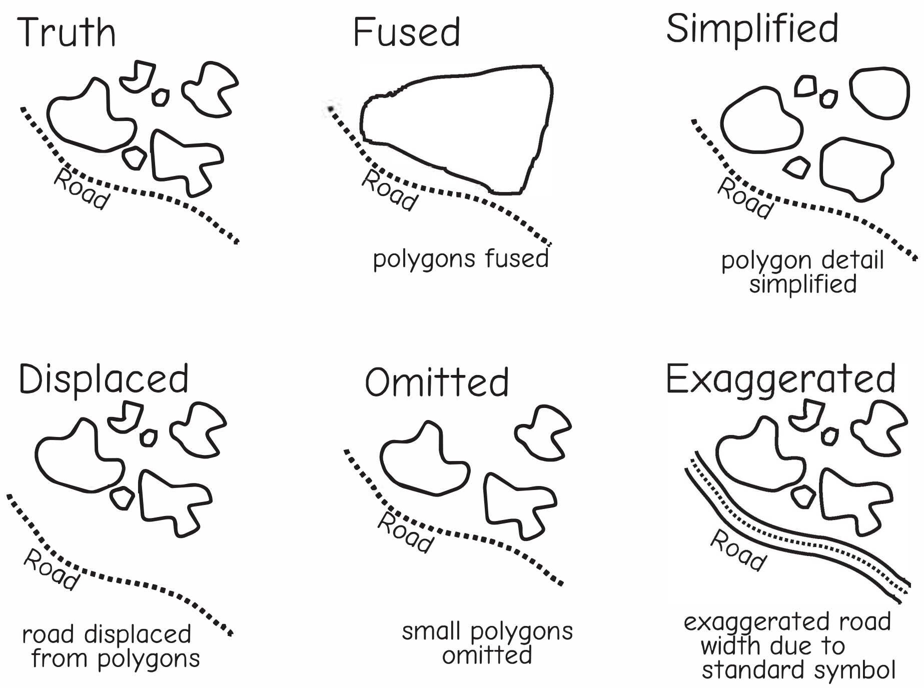
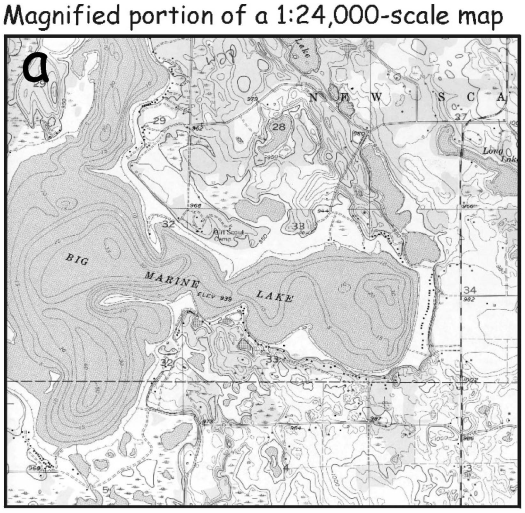
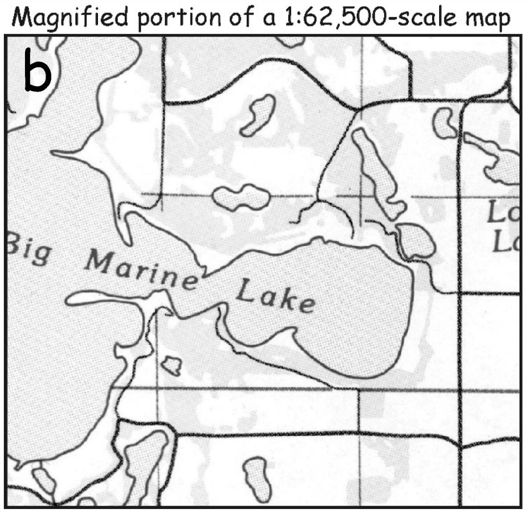
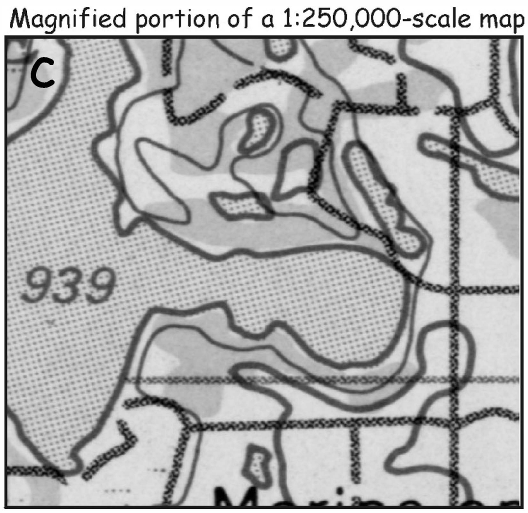


Figure 4-9: Generalizations common in maps and data layers.

Figure 4-10: Examples of map generalization. Portions are shown for three maps for an area in central Minnesota. Excerpts from a large-scale (a, 1:24,000), intermediate-scale (b, 1:62,500), and small-scale (c, 1:250,000) map are shown. **Note that the maps are not drawn at true scale to facilitate comparison**. The smaller-scale maps (b and c) have been magnified more than a to better show the effects of generalization. Each map has a different level of map generalization. Generalizations increase with smaller-scale maps and include omissions of smaller lakes, successively greater road width exaggerations, and increasingly generalized shorelines as one moves from maps a through c.







Map Boundaries and Spatial Data

One final characteristic of maps affects their use as a source of spatial data: digital data are often created or stored in tiles, and hardcopy maps have edges. Discontinuities often occur at tile or map boundaries. These errors may be lacking in many newer, large-area data collected with digital methods, but edges will be encountered and should be understood.

Large-scale, high-quality maps and digital data sets often cover small areas or come broken into smaller tiles. This is because of the trade-off between data volumes, scale, and area coverage, and because of limits on the practical size of a map. High-density image or field-scanned data often tax the storage or display processing capabilities of common computers, and so data are broken into more manageable pieces for access and display. Hardcopy maps above a meter or so in size are expensive and difficult to print, store, or view. Thus, human ergonomics set a practical limit on the physical size of a map.

Because spatial data in a GIS often span large-scale maps or detailed data tiles, these map boundaries may be apparent in a spatial database. Problems often arise when adjacent maps or data layers are entered into a spatial database because features do not align or have mismatched attributes across map boundaries.

Differences in the time of data collection for adjacent map sheets may be a cause of inconsistencies across tile or map borders. Landscape change through time is a major source of differences across map boundaries. For example, scanned and digital aerial photographs are collected for most of the agricultural lands of the U.S. on an annual basis. Data are collected in blocks, over the growing season, and because of weather, schedules, and equipment failures, several days to weeks may pass between data collection on adjacent blocks. Many features, such as crop type, stage of development, wetland size, or harvest state may be discontinuous or inconsistent across blocks, and hence the data layer.

Different interpreters may also cause differences across collection boundaries. Large-area mapping projects typically employ several interpreters, each working on different areas of a region. All professional, large-area mapping efforts should have protocols specifying the scale, sources, equipment, methods, classification, keys, and cross-correlation to ensure consistent mapping across map sheet boundaries. In spite of these efforts, some differences due to human interpretation occur. Feature placement, category assignment, and generalization vary among interpreters. These problems are compounded when extensive checking and guidelines are not enforced across map sheet boundaries, especially when adjacent areas are mapped at different times or by two different organizations.

Finally, differences in coordinate registration can lead to spatial mismatch across map sheets. *Registration* or *coordinate transformation*, discussed later in this chapter, is the conversion of digitizer or other coordinate data to an Earth-surface coordinate system. These registrations contain unavoidable errors that translate into spatial uncertainty. There may be mismatches when data from two separate registrations are joined along the edge of a data layer or map.

Spatial data stored in a GIS are not bound by the same constraints that limit the physical dimensions of hardcopy maps. Digital storage enables the production of seamless digital maps of large areas. However, the inconsistencies that exist in data collection or on hardcopy maps may be transferred to the digital data. Inconsistencies at tile or map sheet edges need to be identified and resolved in digital formats.

Digitizing: Coordinate Capture

Digitizing is the process by which coordinates from a map, image, or other sources are converted into a vector data layer in a GIS. The point, line, and area coordinate values that define the locations and shapes of entities must be captured, that is, recorded as numbers and structured in the spatial database. There is a wealth of spatial data in existing maps and photographs, and new imagery and maps add to this source of information on a nearly continuous basis.

Manual digitization is human-guided coordinate capture from a map or image source. The operator guides an electronic device over a map or image and signals the capture of important coordinates, often by pressing a button on the digitizing device. Important point, line, or area features are traced on the source materials, and the coordinates are recorded in GIS-compatible formats. Valuable data on historical maps may be converted to digital forms through the use of manual digitizing. On-screen digitizing and hardcopy digitizing are the two most common forms of manual digitization.

On-Screen Digitizing

On-screen digitizing, also known as heads-up digitizing, involves manually digitizing on a computer screen, using a digital image as a backdrop. Most image data pro-

duced for spatial data entry since 2010 is delivered with a coordinate system defined for the images, so that data extracted from the images are specified in that coordinate system. Digitizing software allows the operator to trace the points, lines, or polygons that are identified on the image (Figure 4-11) and saves the coordinates and added attribute data into spatial data layers. Digitizing software allows the human operator to specify the type of feature to be recorded, the extent and magnification of the image on screen, the mode of digitizing, and other options to control how data are input. The operator typically guides a cursor over points to be recorded using a mouse, and depresses a button or sequence of buttons to collect the point coordinates. On-screen digitizing can be used for recording information from digital photographs, satellite images, scanned maps, or other image sources.

Hardcopy Map Digitization

Hardcopy digitizing is human-guided coordinate capture from a paper, plastic, or other hardcopy map. An operator securely attaches a map to a digitizing surface and traces lines or points with an electrically sensitized puck (Figure 4-12).The most common digitizers are based on a wire grid embedded in or under a table. Depressing a button specifies the puck location relative to the digitizer coordinate system. Digitizing tables can be quite accurate, with a resolution of between 0.25 and 0.025 mm (0.01 and 0.001 in).

While once a major method for capturing spatial data, hardcopy map digitizing is diminishing in importance as most paper documents have been converted to digital forms. The tables are large, somewhat expensive, and now little used. However, because data from hardcopy sources are

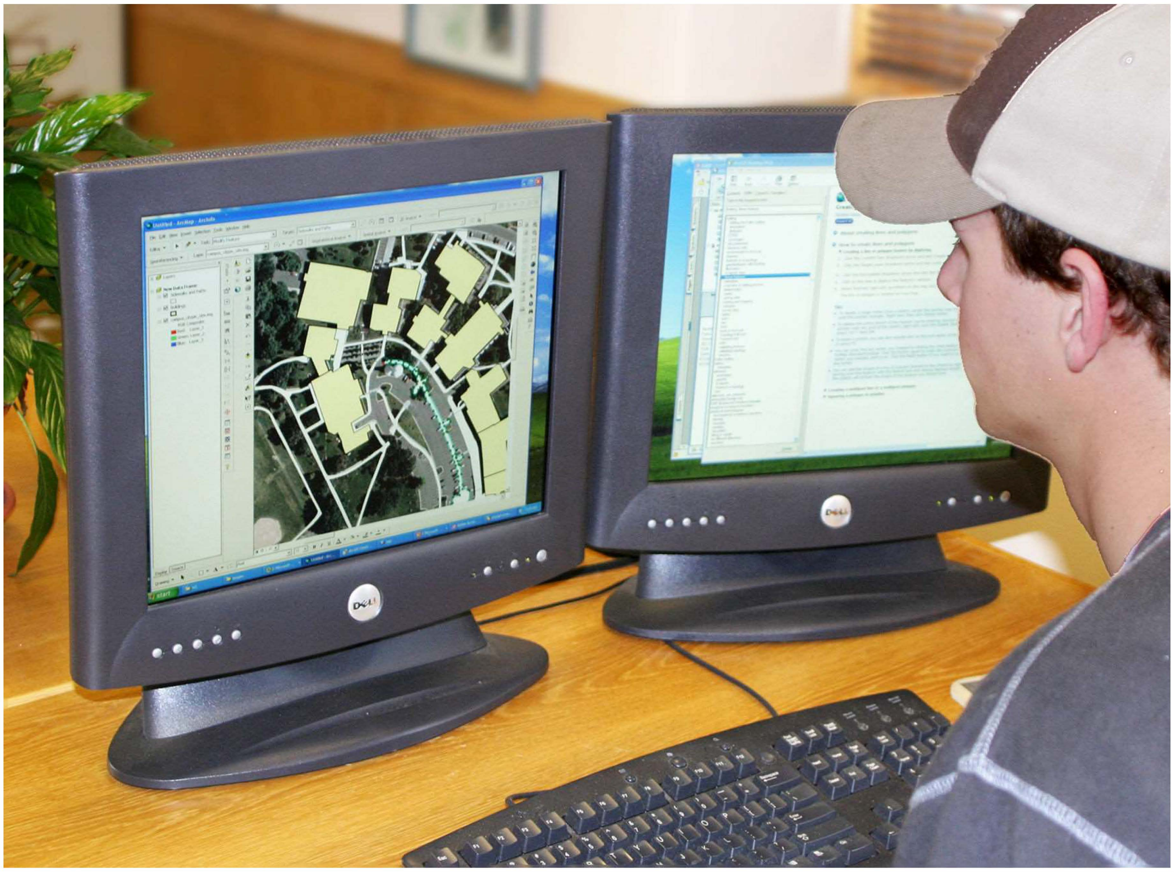


Figure 4-11: An example of on-screen digitizing. Images or maps are displayed on a computer screen and feature data digitized manually. Buildings, roads, or any other features that may be distinguished on the image may be digitized.

likely to persist for many decades, and there are still documents to convert, you should be familiar with the process.

Not all maps are appropriate as a source of information for GIS. The type of map, how it was produced, and the intended purpose must be considered when interpreting the information on maps. Only cartometric maps should be directly digitized, and even though cartometric, a map may not be suitable.

Characteristics of Manual Digitizing

Manual digitizing from digital images or hardcopy sources is common because it provides sufficiently accurate data for most applications. Proper digitizing adds little error to that already in the source materials, requires inexpensive equipment, and is often the best way to provide the needed information. The human ability to interpret images still outpaces that of machines, although artificial intelligence is improving. Humans are usually much better than machines at interpreting the information contained on faded, stained, or poor quality maps and images. Finally, manual digitizing is often best because short training periods are required and data quality may be frequently evaluated. For these reasons, manual digitization is likely to remain an important data entry method for some time to come.

Many factors may affect accuracy while digitizing. As noted earlier, map or image scale and resolution impacts the spatial accuracy of digitized data. This scale may be the production scale for hardcopy maps, or the display scale for digital images or scanned maps. Table 4-1 illustrates the effects of display or scale on data quality. Errors of 1 millimeter (0.039 in) on a 1:24,000 scale map or screen correspond to 24 meters (79 ft) on the surface of the Earth. This same 1 millimeter error on a 1:1,000,000 scale map corresponds to 1,000 meters (3,281 ft) on the Earth's surface. Thus, small errors in image data collection or map production or inter-

Figure 4-12: Manual digitizing on a digitizing table.

pretation may cause significant positional errors when scaled to distances on Earth, and these errors are greater for smaller-scale source materials. Errors due to human pointing ability are reduced for on-screen digitizing, because the operator can zoom in to larger scales as needed. However, this does not overcome errors inherent in original images or scanned documents.

The image source and display scale should be considered when on-screen digitizing, and device precision and map scales should be considered when selecting a hard-copy digitizing. Map scale and repeatability both set an upper limit on the positional quality of digitized data. The most precise digitizers may be required when attempting to meet a stringent error standard while digitizing small-scale maps.

The abilities and attitude of the person digitizing affects the geometric quality of manually digitized data. Operators differ in visual acuity, hand steadiness, attention to detail, and ability to concentrate. The abilities of any single operator will also vary through time due to fatigue or difficulty maintaining focus on a repetitive task. Operators should take frequent breaks from digitizing, and comparisons among operators and quality and consistency checks should be integrated into any manual digitization process to ensure accurate and consistent data collection.

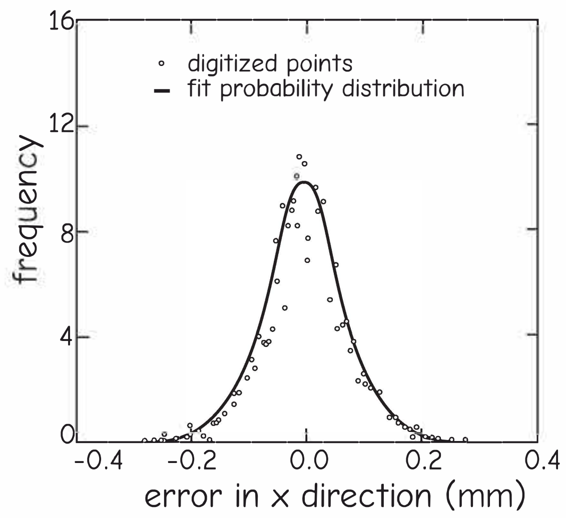


Figure 4-13: Digitizing error, defined by repeat digitizing. Points repeatedly digitized cluster around the true location and follow a normal probability distribution (from Bolstad et al., 1990).

The combined errors from both operators and equipment are usually quite small. One test using a high-precision digitizing table revealed digitizing errors averaging approximately 0.067 mm (Figure 4-13). Errors followed a random normal distribution, and varied significantly among operators. These average errors translated to an approximately 1.6-meter error when scaled from the 1:24,000 scale map to a ground-equivalent distance. This average error is less than the acceptable production error for the map, and is suitable for many spatial analyses.

Table 4-1: The surface error caused by a 1 millimeter (0.039 in) digitizing or map error will change as scale changes. Note the larger error at smaller scales.

Map Scale	Error (m)	Error (ft)
1:24,000	24	79
1:50,000	50	164
1:62,500	63	205
1:100,000	100	328
1:250,000	250	820
1:1,000,000	1,000	3,281

The Digitizing Process

Manual digitizing involves displaying a digital image on screen or placing a map on a digitizing surface, and tracing the location of feature boundaries. Coordinate data are sampled by manually positioning the puck or cursor over each target point and collecting coordinate locations. This position/collect step is repeated for every point to be captured, and all features of interest recorded.

Lines have a *starting node*, then a set of *vertices* defining the line shape, and an *ending node* (Figure 4-14). Hence, lines may be viewed as a series of straight line segments connecting vertices and nodes. Polygons are connected lines that enclose areas.

Digitizing may be in *point mode*, where the operator must depress a button or otherwise signal to the computer to sample each point, or in *stream mode*, where points are automatically sampled at a fixed time or distance frequency, perhaps once each meter. Stream mode helps when large numbers of lines are digitized, because vertices may be sampled more quickly and the operator may become less fatigued.

The stream sampling rate must be specified with care to avoid over- or under-sampled lines. Too short a collection interval results in redundant points. Too long a collection interval may cause the loss of important spatial detail. When using time-triggered stream digitizing, the operator must continuously move the digitizing puck; pausing for a period longer than the sampling interval digitizes multiple points clustered together. These will redundantly represent a line portion, often with overlapping segments. Pausing for extended periods often creates a "rat's nest" of lines that must later be removed.

Minimum distance digitizing is a variant of stream mode digitizing that avoids some of the problems inherent with time-sampled streaming. In minimum distance digitizing, a new point is not recorded unless it is more than some minimum threshold distance from the previously digitized point. The operator may pause without creating a rat's nest of line segments. The threshold must be chosen carefully – neither too large, missing useful detail, nor too small, in effect reverting back to stream digitizing.

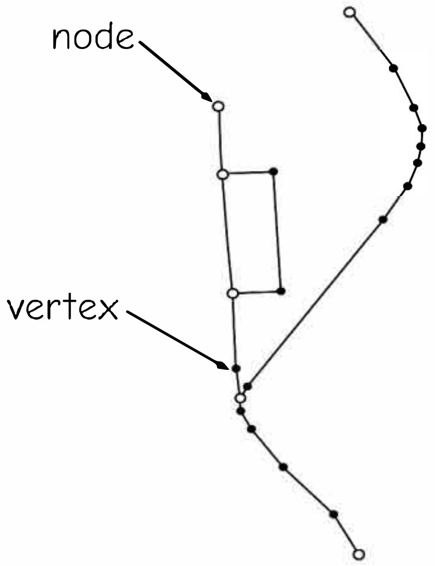


Figure 4-14: Nodes define the starting and ending points of lines; vertices define line shape.

Digitizing Errors, Node and Line Snapping

Positional errors are inevitable when data are manually digitized. These errors may be "small" relative to the intended use of the data; for example, the positional errors may be less than 2 meters when only 5-meter accuracy is required. However, these relatively small errors may still prevent the generation of correct networks or polygons. For example, a data layer representing a river system may not be correct because major tributaries may not connect. Polygon features may not be correctly defined because their boundaries may not completely close. These small errors must be removed or avoided during digitizing. Figure 4-15 shows some common digitizing errors.

Undershoots and *overshoots* are common errors that occur when digitizing. Undershoots are nodes that do not quite reach the line or another node, and overshoots are lines that cross over existing nodes or lines (Figure 4-15). Undershoots

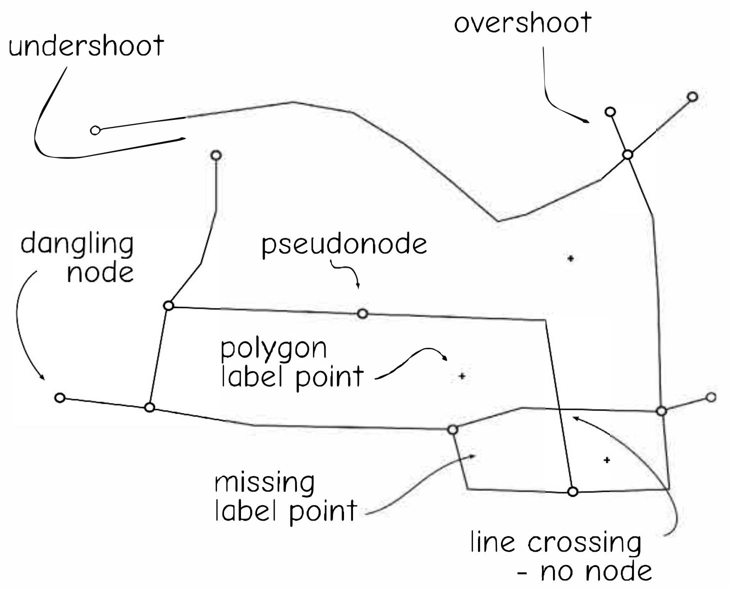


Figure 4-15: Common digitizing errors.

cause unconnected networks and unclosed polygons. Overshoots typically do not cause problems when defining polygons, but they may cause difficulties when defining and analyzing line networks.

Node snapping and *line snapping* are used to reduce undershoots and overshoots while digitizing. Snapping is a process of automatically setting nearby points to have the same coordinates. Snapping relies on a *snap tolerance* or *snap distance*. This distance may be interpreted as a minimum distance between features. Nodes or vertices closer than this distance are moved to occupy the same location (Figure 4-16). While digitizing, an existing node or vertex becomes "magnetic," and pulls a new node or vertex to it within the snap distance. Node snapping prevents a new node from being placed within the snap distance of an already existing node. Remember that nodes are used to define the ending points of a line. By snapping two nodes together, we ensure a connection between digitized lines.

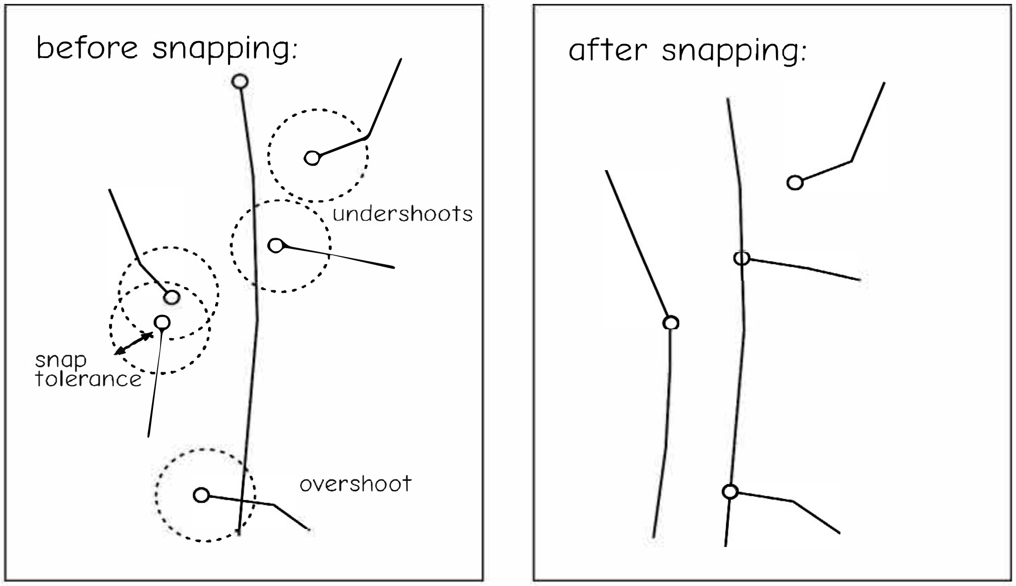


Figure 4-16: Undershoots, overshoots, and snapping. Snapping may join nodes, or may place a node onto a nearby line segment. Snapping does not occur if the nodes and/or lines are separated by more than the snap tolerance.

Line snapping, sometimes called edge snapping, may also be specified. Line snapping inserts a node at a line crossing and clips the end when a small overshoot is digitized. This forces a node to connect to a nearby line while digitizing, but only when the undershoot or overshoot is less than the snapping distance. Line snapping requires the calculation of an intersection point on an already existing line. A snap places a new node at the intersection point, and connects the digitized line to the existing line. This splits the existing line into two new lines. When used properly, line and node snapping reduce the number of undershoots and overshoots. Closed polygons or intersecting lines are easier to digitize accurately and efficiently under node and line snapping.

The snap distance must be carefully selected for snapping. If the snap distance is too short, then snapping has little impact. Consider a configuration where digitized data has better than 5 meter accuracy only 10% of the time. This means 90% of the digitized points will be more than 5-meters from the intended location. If the snap tolerance is set to the equivalent of 0.1 meters, then few points will be snapped. Conversely, we sacrifice accuracy if the snap tolerance is set too large. In our previous example, a snap tolerance of 10 m is higher than our 5 m target, and we may lose significant spatial information contained in the data source. Lines less than 10 meters apart cannot be digitized as separate objects. The snap distance should be smaller than the desired positional accuracy, such that significant detail contained in the digitized map is recorded. It is also important that the snap distance is not below the capabilities of the system used for digitizing. Careful selection of the snap distance may reduce digitizing errors and significantly reduce time required for later editing.

Reshaping: Line Smoothing and Thinning

Digitizing software may provide tools to smooth, densify, or thin points while entering data. One common technique uses *spline* functions to smoothly interpolate curves between digitized points, and thereby both smooth and densify the set of vertices used to represent a line. A spline is a set of polynomial functions that join smoothly (Figure 4-17). Polynomial functions are fit to successive sets of points along the vertices in a line; for example, a function may be fit to points 1 through 5, and a separate polynomial function fit to points 5 through 11 (Figure 4-17). Constraints force these functions to connect smoothly, usually by requiring the first and second derivatives of the functions to be continuous at the intersection point. This means the lines have the same slope at the intersection point, and the slope is changing at the same rate for both lines at the intersection point. Once the spline functions are calculated, they may be used to add vertices. For example, several new vertices may be automatically placed on the line

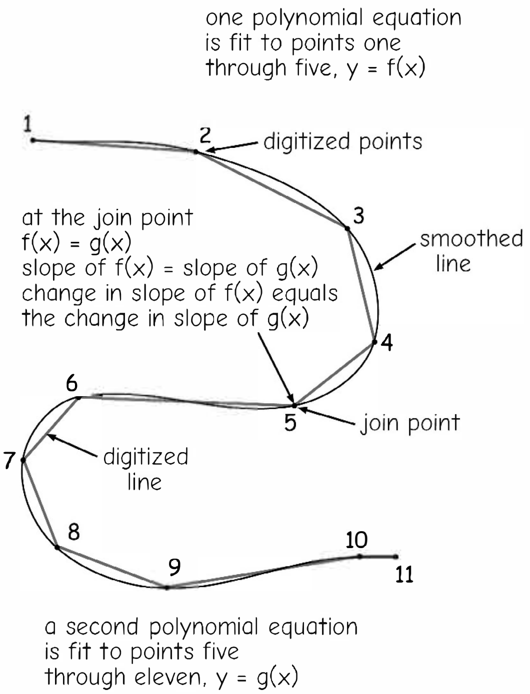


Figure 4-17: Spline interpolation to smooth digitized lines.

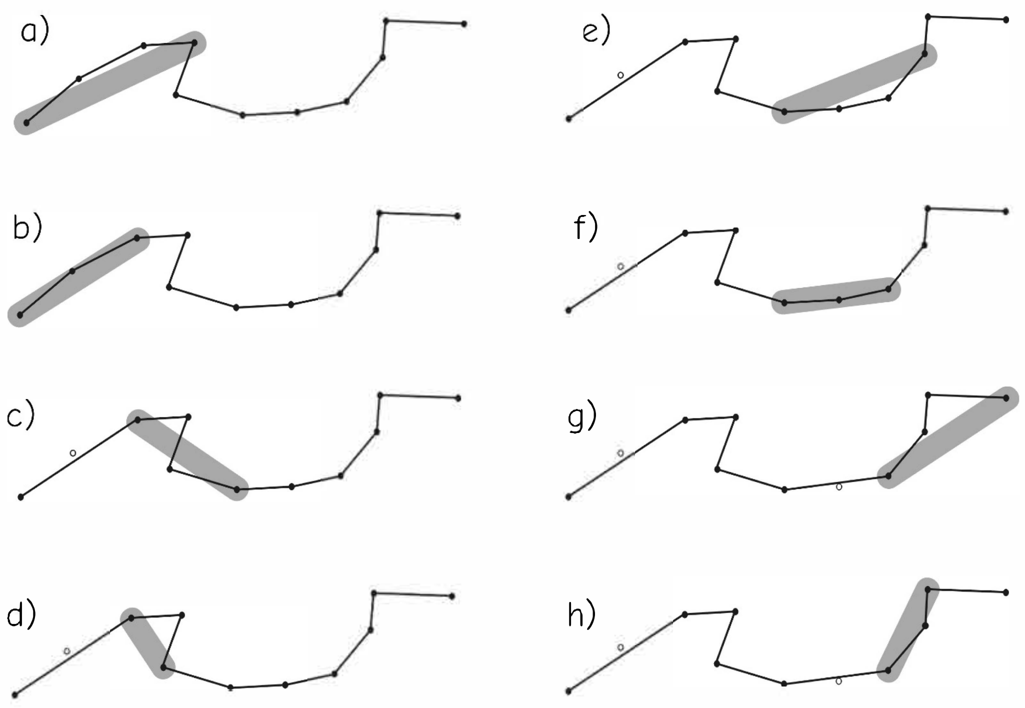


Figure 4-18:The Lang algorithm is a common line-thinning method. In the Lang method, vertices are removed, or thinned, when they are within a weed distance to a spanning line (adapted from Weibel, 1997).

between digitized vertices 8 and 9, leading to the "smooth" curve shown in Figure 4-17.

Data may also be digitized with too many vertices. High densities may occur when data are manually digitized in stream mode, and the operator moves slowly relative to the time interval. High vertex densities may also be found when data are derived from spline or smoothing functions that specify too high a point density. Finally, automated scanning and then raster-to-vector conversion may result in coordinate pairs spaced at absurdly high densities. Many of these coordinate data are redundant and may be removed without sacrificing spatial accuracy. Too many vertices may be a problem in that they slow processing, although this has become less important as computing power has increased. Point thinning algorithms have been developed to reduce the number of points while maintaining the line shape.

Many point thinning methods use a perpendicular "weed" distance, measured from a spanning line, to identify redundant points (Figure 4-18, top). The Lang method exemplifies this approach. A spanning line connects two nonadjacent vertices in a line. A predetermined number of vertices is spanned initially. The initial spanning number has been set to 4 in Figure 4-18, meaning four points will be considered at each starting point. Areas closer than the weed distance are shown in gray in the figure. A straight line is drawn between a starting point and the endpoint that is the fourth point down the line (Figure 4-18a). Any intermediate points that are closer than the weed distance are marked for removal. In Figure 4-18a, no points are within the weed distance, therefore, none are marked. The endpoint is then moved to the next closest remaining point (Figure 4-18b), and all intermediate points tested for removal. Again, any points closer than the weed distance are marked for removal, shown as open circles. Note that in Figure 4-18b, one point is within the weed distance, and is removed. Once all points in the initial spanning distance are checked, the last remaining endpoint becomes the new starting point, and a new spanning line is drawn to connect 4 points (Figure 4-18c, d).

The process may be repeated for successive sets of points in a line segment until all vertices have been evaluated (Figure 4-18e

to h). All close vertices are viewed as not recording a significant change in the line shape, and hence are expendable. Increasing the weed distance thins more vertices, and at some upper weed distance, too many vertices may be removed. A balance must be struck between the removal of redundant vertices and the loss of shape-defining points, usually through a careful set of test cases with successively larger weed distances.

There are many variants on this basic concept. Some look only at three immediately adjacent points, testing the middle point against the line spanned by its two neighboring points. Others constrain or expand the search based on the complexity of the line. Rather than always looking at four points, as in our example above, more points are scrutinized when the line is not complex (nearly straight), and fewer when the line is complex (many changes in direction).

Scan Digitizing

Optical scanning is another method for converting hardcopy documents into digital formats. Scanners have elements that emit and sense light. Most scanners pass a sensing element over an illuminated map. This device measures both the precise location of the point being sensed and the strength of the light reflected or transmitted from that point. Reflected light intensities are converted to numbers, producing a raster representation of the map. Values are recorded where points or lines exist on the map, and null or zero values are recorded in the intervening spaces.

Scan digitization usually requires some form of *skeletonizing*, or line thinning, particularly if the data are to be converted to a vector data format. Scanned lines are often wider than a single pixel (Figure 4-19). One of several pixels may be selected to specify the position of a given portion of the line. The same holds true for points. A pixel near the "center" of the point or line is typically chosen, with the center of a line defined as the pixel nearest the center of the local perpendicular bisector of the line. Skeletonizing reduces the widths of lines or points to a single pixel.

Editing Geographic Data

Spatial data may be edited, or changed, for several reasons. Errors and inconsistencies are inevitably introduced during spatial data entry. Undershoots, overshoots, missing or extra lines, and missing or extra points or labels are all errors that must be corrected. Spatial data can change over time. Parcels are subdivided, roads extended or moved, forests grow or are cut, and these changes may be entered in the spatial database through editing. New technologies may be

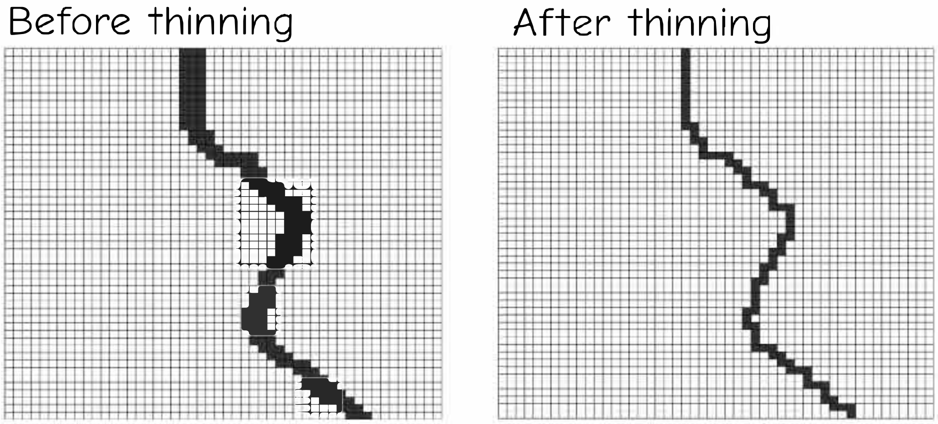


Figure 4-19: Skeletonizing, a form of line thinning that is often applied after scan digitizing.

developed that provide more accurate positional information, and even though existing data may be consistent and current, the more accurate data may be more useful, leading to data editing.

Software helps operators identify potential errors. Line nodes may be classified as connecting or dangling. A connecting node joins two or more lines, while a dangling node is attached to only one line. Some dangling nodes may be intentional, for example, a cul-de-sac in a street network, while others will be the result of under- or overshoots. Dangling nodes can be quickly evaluated and, if appropriate, corrected.

Attribute consistency may also be used to identify errors. Operators note areas in which contradictory theme types occur in different data layers. The two layers are either graphically or cartographically overlain. Contradictory co-occurrences are identified, such as water in one layer and upland areas in a second. These contradictions are then either resolved manually, or automatically via some predefined precedence hierarchy.

Many GIS software packages provide a comprehensive set of editing tools (Figure 4-20). Editing typically includes the ability to select, split, update, and add features. Selection may be based on geometric attributes, or with a cursor guided by the operator. Selections may be made individually, by geographic extent (select all features in a box, circle, or within a certain distance of the pointer), or by geometric attributes (e.g., select all nodes that connect to only one line). Once a feature is selected, various operations may be available, including eras-

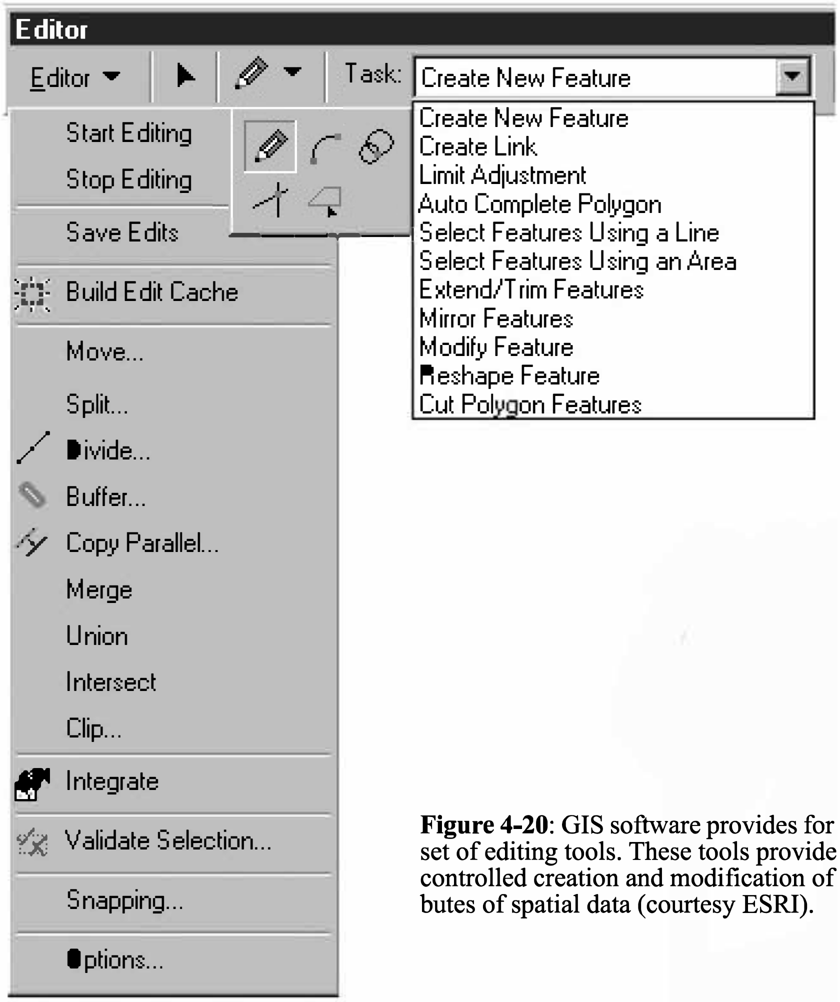


Figure 4-20: GIS software provides for a flexible and complete set of editing tools. These tools provide for the rapid, precise, controlled creation and modification of coordinates and attributes of spatial data (courtesy ESRI).

ing all or part of the feature, changing the coordinate values defining the feature, and in the case of lines, splitting or adding to the feature. A line may be split into parts, either to isolate a segment for future deletion, or to modify only a portion of the line. Coordinates are typically altered by interactively selecting and dragging points, nodes, or vertices to their best shape and location. Points or line segments are added as needed.

Groups of features in an area may be adjusted through interactive *rubbersheeting*. Rubbersheeting involves fitting a local equation to adjust the coordinates of features. Polynomial equations are often used due to their flexibility and ease of application. Anchor points are selected, again on the graphics screen, and other points are selected by dragging interactively on the screen to match point locations. All lines and points except the anchor points are interactively adjusted.

All edits should be made with due attention to distance shifts during editing. On-screen editing to eliminate undershoots should only be performed when the "true" locations of features may be identified accurately, and the new features can be confidently placed in the correct location. Automatic removal of "short" undershoots may be performed without introducing additional spatial error in most instances. A short distance for an undershoot is subjectively defined, but typically it is below the error inherent in the source map, or at least a distance that is insignificant when considering the intended use of the spatial data.

Features Common to Several Layers

One common problem in digitizing derives from representation of features that occur on different maps or images. These features rarely have identical locations on each map or image, and often occur in different locations when digitized into their respective data layers (Figure 4-21). For example, water boundaries on soil survey maps rarely correspond exactly to water boundaries found on USGS topographic maps.

Features may appear differently on different maps for many reasons. Perhaps the maps were made for different purposes or at different times. Features may differ because the maps were from different source materials; for example, one map may have been based on ground surveys while another was based on aerial photographs. Digitizing can also compound the problem due to differences in digitizing methods or operators.

There are several ways to remove this "common feature" inconsistency. One involves removing inconsistencies while redrafting the data from conflicting sources onto a new base map. Redrafting is labor intensive and time consuming, but forces a resolution of inconsistent boundary locations. Redrafting also allows several maps to be combined into a single data layer.

A second, often preferable method involves establishing a "master" boundary that is the highest accuracy composite of the available data sets. A digital copy or overlay operation establishes the common features as a base in all the data layers, and this base may be used as each new layer is produced. For example, water boundaries might be extracted from the soil survey and USGS quadrangle maps, and these data combined in a third data layer. The third data layer would be edited to produce a composite, high-quality water layer. The composite water layer would then be copied back into both the soils and USGS quad layers. This second approach, while resulting in visually consistent spatial data layers, is in many

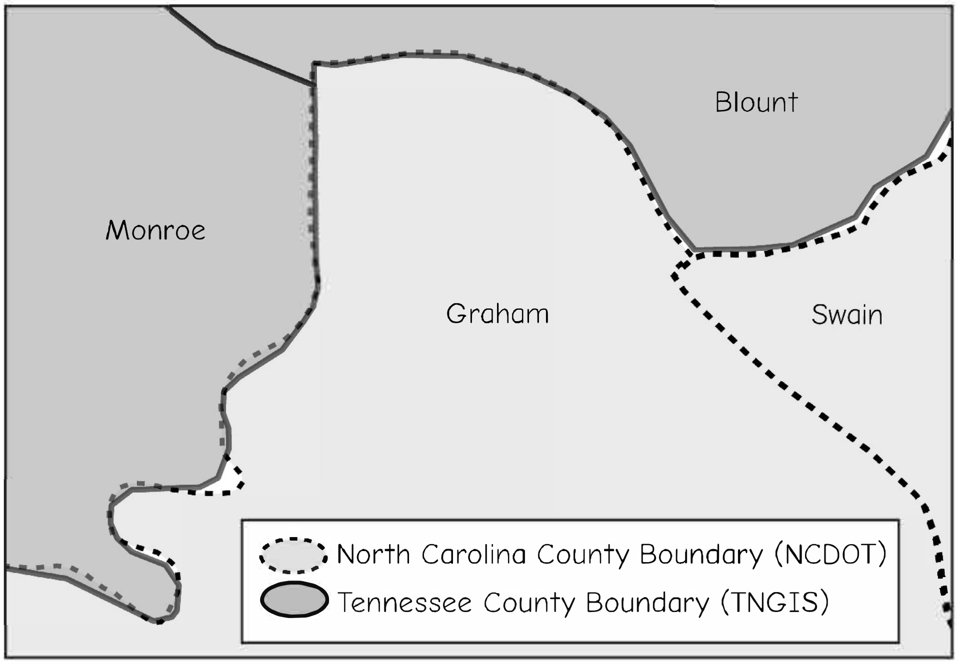


Figure 4-21: Common features may be spatially inconsistent in different spatial data layers. These data were the highest resolution data available from the respective official state distribution nodes. Note the boundaries are inconsistent in several locations, creating gaps and overlaps along their shared margin. These gaps must be addressed if the data sets are to be combined in an analysis.

instances only a cosmetic improvement of the data. If there are large discrepancies ("large" is defined relative to the required spatial data accuracy), then the source of the discrepancies should be identified and the most accurate data used, or new, higher-accuracy data collected from the field or original sources.

Coordinate Transformation

Coordinate transformation is a common operation in the development of spatial data for GIS. A coordinate transformation brings spatial data into an Earth-based map coordinate system so that each data layer aligns with every other data layer. This alignment ensures features fall in their proper relative position when digital data from different layers are combined. Within the limits of data accuracy, a good transformation helps avoid inconsistent spatial relationships such as farm fields on freeways, roads under water, or cities in the middle of swamps, except where these truly exist. Coordinate transformation is also referred to as *registration*, because it “registers” the layers to a map coordinate system.

Coordinate transformation is most commonly used to convert newly digitized data from the digitizer/scanner coordinate system to a standard map coordinate system (Figure 4-22). The input coordinate system is usually based on the digitizer or scanner-assigned values. An image may be scanned and coordinates recorded as a cursor is moved across the image surface. These coordinates are usually recorded in pixel, inch, or centimeter units relative to an origin located near the lower left corner of the image. The absolute values of the coordinates depend on where the image happened to be placed on the table prior to scanning, but the relative position of digitized points does not change. Before these newly digitized data may be used with other data, these “inch-space” or “digitizer” coordinates must be transformed into an Earth-based map coordinate system.

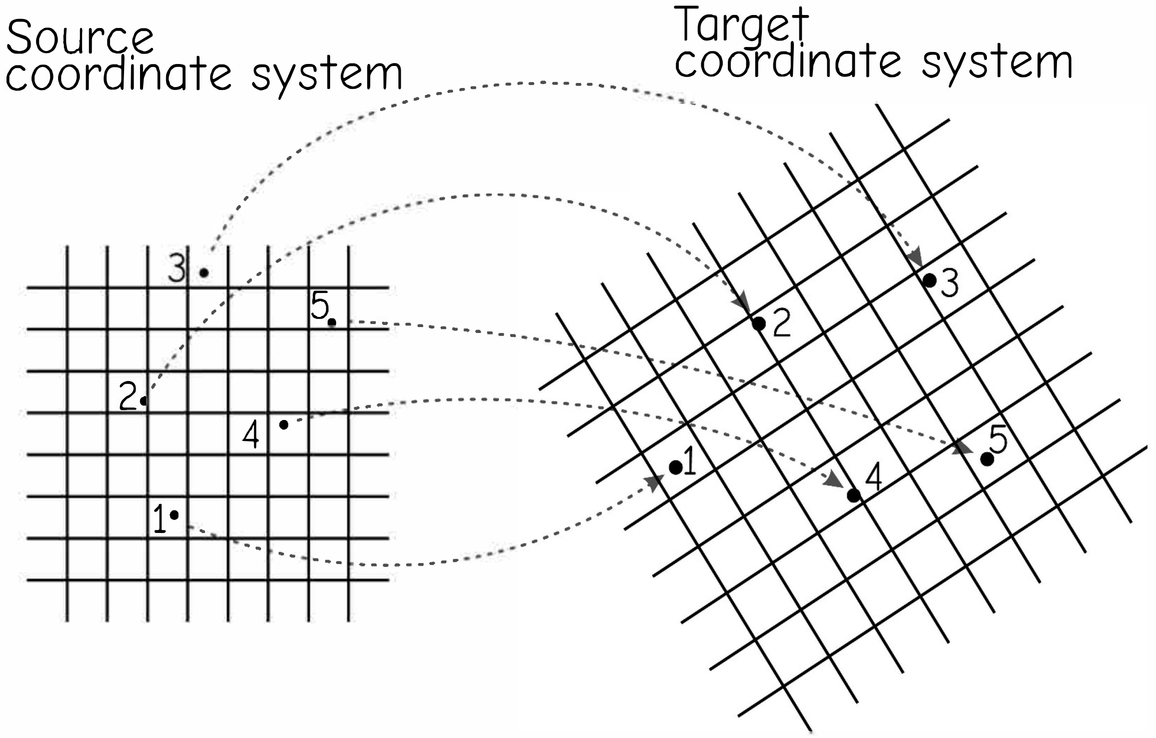


Figure 4-22: Control points in a coordinate transformation. Control points are used to guide the transformation of a source, input set of coordinates to a target, output set of coordinates. There are five control points in this example. Corresponding positions are shown in both coordinate systems.

Figure 4-23 depicts the application of a coordinate transformation in data development. Early surveys were often stored on paper maps; in this instance, the original PLSS surveys. These depict the original PLSS boundary lines, as well as lakes and wetlands, and in some instances forests, grasslands, or other features. We may wish to compare past and current conditions, but the PLSS is a system of land division, with no coordinates or projection associated with the lines. We need to register the paper maps to a projected coordinate system prior to use.

The PLSS line intersections may be used to register the original maps to a projected coordinate system. As noted in Chapter 3, PLSS lines often became property boundaries, and subsequent roads often followed these boundaries. A road between properties shared the imposition among property owners, and served multiple adjacent properties. Road intersections often occur at PLSS corners, particularly in the flatter terrain over much of the eastern and central United States. Section line intersections may be surveyed directly, or extracted

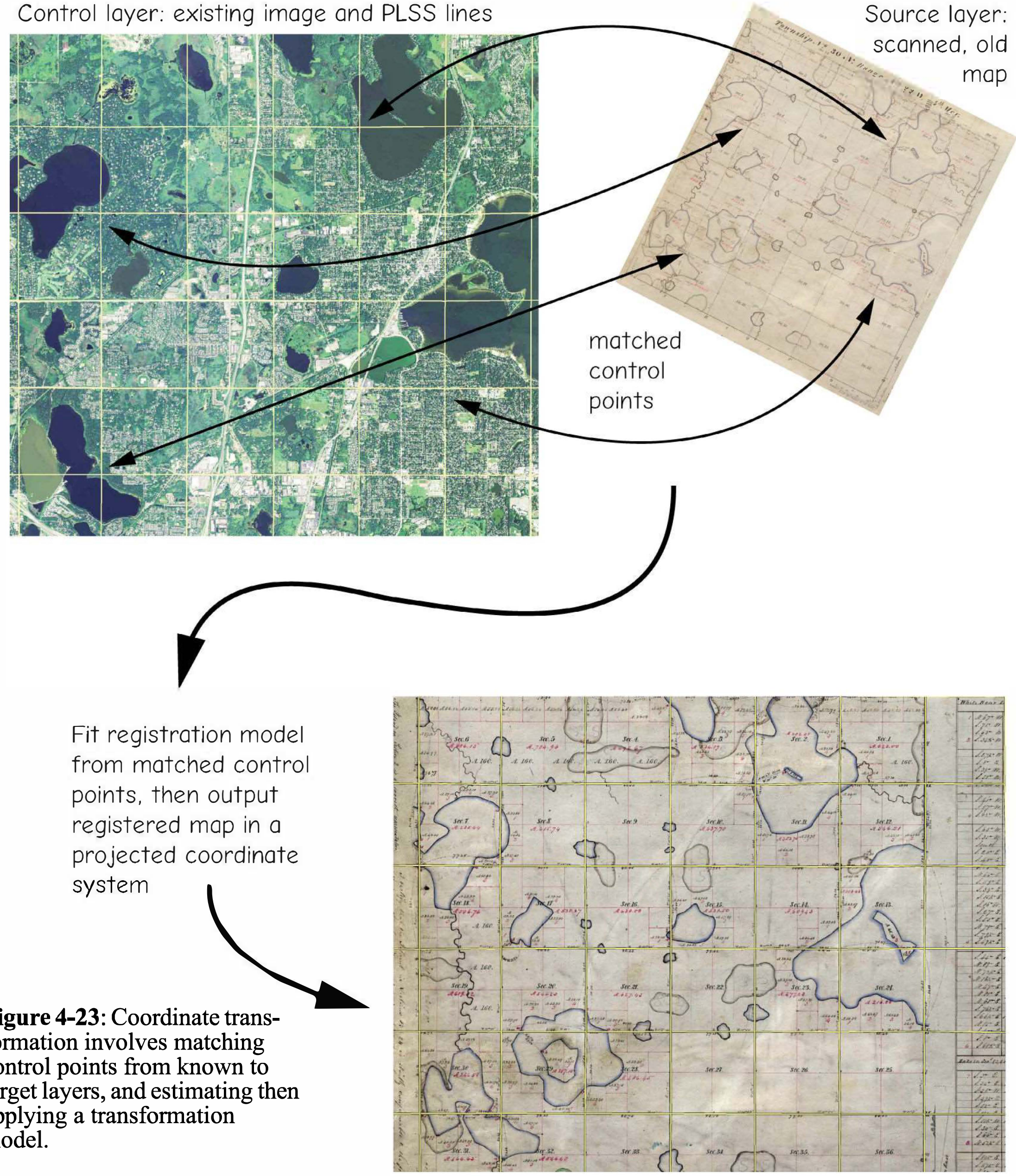


Figure 4-23: Coordinate transformation involves matching control points from known to target layers, and estimating then applying a transformation model.

from other registered data such as aerial images, and hence used to transform the original maps to a projected coordinate system. Features on the original maps, such as lakes or wetlands depicted at the time of the original surveys during the 1800s, may be digitized and compared to current feature locations.

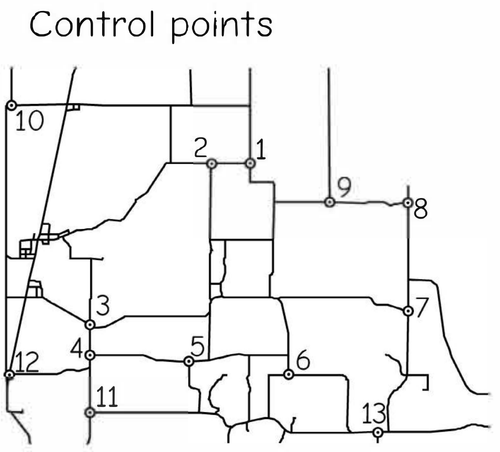
Control Points

A set of *control points* is used to transform the digitized data from the digitizer or photo coordinate system to a map-projected coordinate system. These control points are used to estimate equations that we use for the coordinate transformation (Figure 4-24). Control points are different from other digitized features. When we digitize most points, lines, or areas, we do not know the map projection coordinates for these features. We simply collect the digitizer X and Y coordinates that are established with reference to some arbitrary origin on an image or digitizing table. Control points differ from other digitized points in that we know both the map projection coordinates and the digitizer coordinates for these points.

These two sets of coordinates for each control point, one for the map projection and one for the digitizer system, are used to estimate the coefficients for transformation equations, usually through a statistical, least squares process. The transformation equations are then used to convert coordinates from the digitizer system to the map projection system.

The transformation may be estimated in the initial digitizing steps, and applied as the coordinates are digitized from the map or image. This "on-the-fly" transformation allows data to be output and analyzed with reference to map-projected coordinates. A previously registered data layer or image may be displayed on screen just prior to digitizing a new map. Control points may then be entered, the new map attached to the digitizing table, and the map registered. The new data may then be displayed on top of the previously registered data. This allows a quick check on the location of the newly digitized objects against corresponding objects in the study area.

In contrast to on-the-fly transformations, data can also be recorded in digitizer coordinates, and the transformation applied later. All data are digitized, including the control point locations. The digitizer coordinates of the control point may then be matched to corresponding map projection coordinates, and transformation equations estimated. These transformation equations are then



Digitizer coordinates			Projection coordinates (UTM)	
ID	X	Y	E	N
1	103.0	-100.1	500,083.4	5,003,683.5
2	0.8	-69.1	504,092.3	5,002,499.5
3	-20.0	-69.0	504,907.5	5,002,499.5
4	-60.0	-47.0	506,493.3	5,001,673.5
5	-102.0	-47.2	508,101.3	5,001,651.7
6	-101.7	10.8	508,090.1	4,999,384.0
7	-86.0	75.8	507,475.9	4,996,849.0
8	-40.0	45.7	505,689.2	4,998,022.0
9	11.0	36.8	503,679.2	4,998,368.0
10	63.0	34.0	501,657.9	4,998,479.5
11	63.0	17.7	501,669.1	4,999,116.0
12	63.0	64.3	501,680.3	4,997,296.0
13	106.0	47.7	500,005.3	4,997,943.5

Figure 4-24: An example of control point locations from a road data layer, and corresponding digitizer and map projection coordinates.

applied to convert all digitized data to map projection coordinates.

Control points should meet several criteria. First, they should be from a source with the highest feasible accuracy. Second, control point should be more accurate than the desired overall positional accuracy for the spatial data. Third, control points should be evenly distributed throughout the data area. A sufficient number of control points should be collected, above the minimum to improve the statistically fit transformation functions.

The x, y (horizontal), and sometimes z (vertical or elevation) coordinates of control points are known to a high degree of accuracy and precision. Because high precision and accuracy are subjectively defined, there are many methods to determine control point locations. Subcentimeter accuracy may be required for control points used in property boundary layers, while accuracies of a few meters may be acceptable for large-area vegetation mapping. Common sources of control point coordinates are traditional transit and distance surveys, global positioning system measurements, existing cartometric quality maps, or existing digital data layers on which suitable features may be identified.

The Affine Transformation

The *affine coordinate transformation* employs linear equations to calculate map coordinates. Map projection coordinates are often referred to as eastings (E) and northings (N), and are related to the x and y digitizer coordinates by the equations:

$$E = T_E + a_1 x + a_2 y \qquad (4.1)$$

$$N = T_N + b_1 x + b_2 y \qquad (4.2)$$

Equations 4.1 and 4.2 allow us to move from the arbitrary digitizer coordinate system to the project map coordinate system. We know the x and y coordinates for every digitized point, line vertex, or polygon vertex. We may calculate the E and N coordinates by applying the above equations to every digitized point.

T_E and T_N can be thought of as shifts, or translations, in the origins from one coordinate system to the next. The a_i and b_i parameters incorporate the change in scales and rotation angle between coordinate systems. . The affine is the most commonly applied coordinate transformation because it provides for these three main effects of translation, rotation, and scaling, and because it often introduces less error than higher-order polynomial transformations.

The affine system of equations has six parameters to be estimated: T_E, T_N, a_1, a_2, b_1, and b_2. Each control point provides E, N, x, and y coordinates, and allows us to write two equations. For example, we may have a control point consisting of a precisely surveyed center of a road intersection. This point has digitizer coordinates of x = 103.0 centimeters and y = -100.1 centimeters, and corresponding Earth-based map projection coordinates of E = 500,083.4 and N = 4,903,683.5. We may then write two equations based on this control point:

$$500{,}083.4 = T_E + a_1(103.0) + a_2(-100.1) \qquad (4.3)$$

$$4{,}903{,}683.5 = T_N + b_1(103.0) + b_2(-100.1) \qquad (4.4)$$

We cannot find a unique solution to these equations, because there are six unknowns (T_E, T_N, a_1, a_2, b_1, b_2) and only two equations. We need as many equations as unknowns to solve a linear system of equations. Each control point gives us two equations, so we need a minimum of three control points to estimate the parameters of an affine transformation. Statistical estimation requires a total of four control points. As with all statistical estimates, more control points are better than fewer, but we will

reach a point of diminishing returns after some number of points, typically somewhere between 18 and 30 control points.

The affine coordinate transformation is usually fit using a statistical method that minimizes the *root mean square error* (RMSE). The RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{e_1^2 + e_2^2 + e_3^2 \ldots + e_n^2}{n}} \quad (4.5)$$

where the e_i are the residual distances between the true E and N coordinates and the E and N coordinates in the output data layer:

$$e = \sqrt{(x_t - x_d)^2 + (y_t - y_d)^2} \quad (4.6)$$

This residual is the difference between the true coordinates x_t, y_t, and the transformed output coordinates x_d, y_d. Figure 4-25 shows examples of this lack of fit. Individual residuals may be observed at each control point location.

A statistical method for estimating transformation equations is preferred

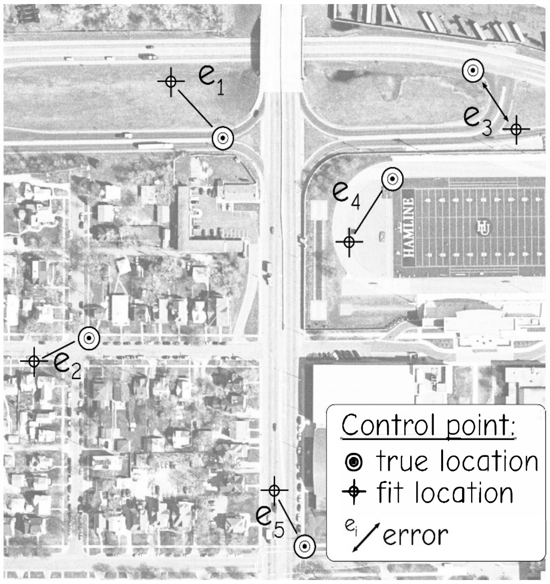


Figure 4-25: Examples of control points, predicted control locations, and residuals from coordinate transformation.

because it also identifies transformation error. Control point coordinates contain unavoidable measurement errors. A statistical process provides an RMSE, a summary of the differences between the "true" (measured) and predicted control point coordinates. It provides one index of transformation quality. Transformations are fit (Figure 4-26). The RMSE will usually be less than the true transformation error at a randomly selected point, because we are actively minimizing the N and E residual errors when we statistically fit the transformation equations. However, the RMSE is an index of accuracy, and a lower RMSE generally indicates a more accurate affine transformation.

Estimating the coordinate transformation parameters is often an iterative process. Control points are rarely exact, and x and y coordinates may not be precisely digitized. Poor eyesight, a shaky hand, fatigue, lack of attention, misidentification of the control location, or a blunder may result in erroneous x and y values. Typically, control points are entered, the affine transformation parameters estimated, and the overall RMSE and individual point E and N errors evaluated (Figure 4-25, Figure 4-26). Suspect points are fixed, and the transformation re-estimated and errors evaluated until a final transformation is estimated. The transformation is then applied to all features to convert them from digitizer to map coordinates.

Model Fit 1:

E = 1.3325289 * x
+ 0.0058654 * y
- 206851.8

N = - 0.002886 * x
+ 1.3296931 * y
- 1660286

RMSE = 9.36

Examine points 15 & 17, adjust noted blunders, refit model

1	518,687.6	5,015,347.0	513,734.1	5,007,087.4	3.07
2	516,907.3	5,013,549.1	511,355.8	5,004,707.2	8.13
3	516,952.2	5,017,965.3	511,438.3	5,010,573.9	4.38
4	518,700.1	5,014,393.4	513,738.9	5,005,831.3	10.99
5	518,099.6	5,013,576.2	512,938.9	5,004,733.6	2.79
6	518,992.6	5,017,306.0	514,144.0	5,009,699.3	8.18
7	519,150.0	5,013,556.6	514,331.9	5,004,709.3	5.66
8	519,259.8	5,013,600.0	514,482.8	5,004,764.0	0.88
9	516,916.8	5,016,528.9	511,378.9	5,008,669.6	4.05
10	516,659.6	5,018,093.8	511,043.8	5,010,744.1	3.37
11	519,474.3	5,018,046.9	514,807.0	5,010,675.2	11.05
12	519,549.2	5,014,375.9	514,873.0	5,005,798.0	2.84
13	518,089.4	5,014,478.2	512,938.6	5,005,931.0	10.36
14	518,087.4	5,014,755.2	512,936.0	5,006,299.0	9.16
15	518,079.9	5,016,484.0	512,912.3	5,008,591.6	19.49
16	516,947.5	5,017,736.1	511,424.5	5,010,277.6	7.16
17	517,016.3	5,014,444.0	511,485.6	5,005,903.6	17.05
18	517,785.1	5,017,492.6	512,542.4	5,009,954.0	9.51
19	519,435.7	5,017,340.7	514,736.0	5,009,735.7	4.46
20	518,710.3	5,016,544.2	513,778.7	5,008,679.7	10.04
21	518,984.0	5,016,548.6	514,127.8	5,008,678.2	9.50
22	516,719.0	5,014,555.9	511,106.7	5,006,028.1	14.96

Model Fit 2:

E = 1.3319386 * x
+ 0.0057193 * y
- 205812.1

N = - 0.002462 * x
+ 1.329962 * y
- 1161855

RMSE = 7.72

Examine points 4 & 22, adjust noted blunders, refit model

1	518,687.6	5,015,347.0	513,734.1	5,007,087.4	2.56
2	516,907.3	5,013,549.1	511,355.8	5,004,707.2	7.22
3	516,952.2	5,017,965.3	511,438.3	5,010,573.9	3.21
4	518,700.1	5,014,393.4	513,738.9	5,005,831.3	11.45
5	518,099.6	5,013,576.2	512,938.9	5,004,733.6	1.77
6	518,992.6	5,017,306.0	514,144.0	5,009,699.3	7.79
7	519,150.0	5,013,556.6	514,331.9	5,004,709.3	6.34
8	519,259.8	5,013,600.0	514,482.8	5,004,764.0	1.38
9	516,916.8	5,016,528.9	511,378.9	5,008,669.6	4.62
10	516,659.6	5,018,093.8	511,043.8	5,010,744.1	4.09
11	519,474.3	5,018,046.9	514,807.0	5,010,675.2	11.90
12	519,549.2	5,014,375.9	514,873.0	5,005,798.0	2.79
13	518,089.4	5,014,478.2	512,938.6	5,005,931.0	9.15
14	518,087.4	5,014,755.2	512,936.0	5,006,299.0	8.04
15	518,079.1	5,016,483.3	512,921.1	5,008,596.5	7.71
16	516,947.5	5,017,736.1	511,424.5	5,010,277.6	9.22
17	517,015.8	5,014,443.1	511,495.1	5,005,894.9	6.36
18	517,785.1	5,017,492.6	512,542.4	5,009,954.0	9.62
19	519,435.7	5,017,340.7	514,736.0	5,009,735.7	4.88
20	518,710.3	5,016,544.2	513,778.7	5,008,679.7	8.78
21	518,984.0	5,016,548.6	514,127.8	5,008,678.2	10.08
22	516,719.0	5,014,555.9	511,106.7	5,006,028.1	13.68

Model Fit 3:

E = 1.33118637 * x
+ 0.0056629 * y
- 205490.3

N = - 0.003516 * x
+ 1.3297296 * y
- 1160143

RMSE = 6.78

Examine points, no more blunders found.

1	518,687.6	5,015,347.0	513,734.1	5,007,087.4	2.48
2	516,907.3	5,013,549.1	511,355.8	5,004,707.2	5.63
3	516,952.2	5,017,965.3	511,438.3	5,010,573.9	3.84
4	518,699.6	5,014,396.8	513,739.3	5,005,831.0	6.62
5	518,099.6	5,013,576.2	512,938.9	5,004,733.6	2.71
6	518,992.6	5,017,306.0	514,144.0	5,009,699.3	8.40
7	519,150.0	5,013,556.6	514,331.9	5,004,709.3	6.55
8	519,259.8	5,013,600.0	514,482.8	5,004,764.0	1.52
9	516,916.8	5,016,528.9	511,378.9	5,008,669.6	3.30
10	516,659.6	5,018,093.8	511,043.8	5,010,744.1	5.27
11	519,474.3	5,018,046.9	514,807.0	5,010,675.2	11.78
12	519,549.2	5,014,375.9	514,873.0	5,005,798.0	3.54
13	518,089.4	5,014,478.2	512,938.6	5,005,931.0	9.34
14	518,087.4	5,014,755.2	512,936.0	5,006,299.0	8.30
15	518,079.1	5,016,483.3	512,921.1	5,008,596.5	5.67
16	516,947.5	5,017,736.1	511,424.5	5,010,277.6	6.73
17	517,015.8	5,014,443.1	511,495.1	5,005,894.9	3.30
18	517,785.1	5,017,492.6	512,542.4	5,009,954.0	8.86
19	519,435.7	5,017,340.7	514,736.0	5,009,735.7	4.13
20	518,710.3	5,016,544.2	513,778.7	5,008,679.7	9.55
21	518,984.0	5,016,548.6	514,127.8	5,008,678.2	9.53
22	516,717.8	5,014,546.4	511,106.4	5,006,028.8	9.01

Figure 4-26: Iterative fitting of an affine transformation. Control points were examined after each fit, to discover blunders in entry or poor matching of points. Control points with large residuals were examined to determine if the cause for the error may be identified. If so, the control point coordinates may be modified, and transformation refit.

Other Coordinate Transformations

Other coordinate transformations are sometimes used. The *conformal coordinate transformation* is similar to the affine, and has the form:

$$E = T_E + cx - dy \quad (4.7)$$

$$N = T_N + dx + cy \quad (4.8)$$

The coefficients T_E, T_N, c, and d are estimated from control point data. Like the affine transformation, the conformal transformation is also a first-order polynomial. Unlike the affine, the conformal transformation requires equal scale changes in the x and y directions. Note the symmetry in equations 4.7 and 4.8, in that the x and y coefficients match across equations, and there is a change in sign for the d coefficient. This results in a system of equations with only four unknown parameters, and so the conformal may be estimated when only two control points are available.

Higher-order polynomial transformations are sometimes used to transform among coordinate systems. An example of a 2nd-order polynomial is:

$$E = b_1 + b_2x + b_3y + b_4x^2 + b_5y^2 + b_6xy \quad (4.9)$$

Note that the combined powers of the x and y variables may be up to 2. This allows for curvature in the transformation in both the x and y directions. A minimum of six control points is required to fit this second-order polynomial transformation, and seven are required when using a statistical fit. The estimated parameters T_E, T_N, a_1, a_2, b_1, and b_2 will be different in equations 4.1 and 4.2 when compared to 4.9, even if the same set of control points is used for both statistical fits. We change the form of the equations by including the higher-order squared and xy cross-product terms, and all estimated parameters will vary.

A Caution When Evaluating Transformations

Selecting the "best" coordinate transformation to apply is a subjective process, guided by multiple goals. We hope to develop an accurate transformation based on a large set of well-distributed control points. Isolated control points that substantially improve our coverage may also contribute substantially to our transformation error.

There are no clear rules on the number of points versus distribution of points trade-off, but it is typically best to strive for the widest distribution of points. We want at least two control points in each quadrant of the working area, with a target of 20% in each quadrant. These goals are often not impossible. The transformation equation should be developed with the following observations in mind.

First, bad control points happen, but we should thoroughly justify the removal of any control point. Every attempt should be made to identify the source of the error, either in the collection or in the processing of field coordinates, the collection of image coordinates, or in some blunder in coordinate transcription. A common error is the misidentification of coordinate location on the image or map, for example, when the control location is placed on the wrong side of a road.

Second, a lower RMSE does not mean a better transformation. The RMSE is a useful tool when comparing among transformations that have the same model form, for example, when comparing one affine to another affine, as in Figure 4-26. The RMSE is not useful when comparing among different model forms, for example, when comparing an affine to a second-order polynomial. The RMSE is typically lower for a second- and other higher-order polynomials than an affine transformation, but this does not mean

the higher-order polynomial provides a more accurate transformation. The higher-order polynomial will introduce more error than an affine transformation on most orthographic maps, and an affine transformation is preferred. High-order polynomials allow more flexibility in warping the surface to fit the control points. Unfortunately, this warping may significantly deform the non-control-point coordinates, and add large errors when the transformation is applied to all data in a layer (Figure 4-27). Thus, high-order polynomials and others should be used with caution.

Finally, independent tests of the transformations make the best comparisons among transformations. A completely independent set of widely distributed test points are ideal, but these rarely exist. The extra points either haven't been collected, or suitable locations do not exist. The best way to test the accuracy of the transformation typically uses a "bootstrap" approach that treats each point as an independent test point. One point is withheld, the transformation estimated, and the error at the withheld point calculated. The point is replaced in the estimation set, and the next point withheld, fitting the same type of transformation. The equations will be slightly different. The error at this second withheld point is then calculated. This process is repeated for each control point, and a mean error calculated.

Control Point Sources: Surveying

Traditional ground surveys based on optical surface measurements are a common, although decreasingly used method for determining control point locations. Federal, state, county, and local governments all maintain a set of accurately surveyed locations (Figure 4-28), and these points may be used as control points or as starting points for additional surveys. Many of these known

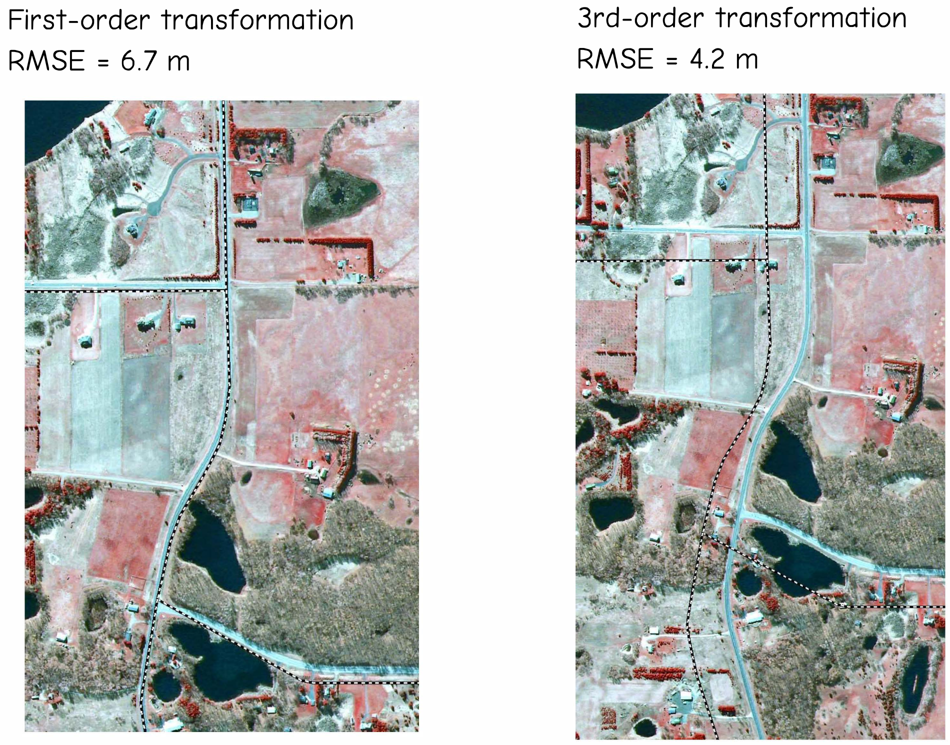


Figure 4-27: An illustration that RMSE should not be used to compare different order transformations, nor should it be used as the sole criterion for selecting the best transformation. Above are portions of a transformed image that was registered to a road network. This area is interstitial to 18 widely distributed control points. Because the 3rd-order polynomial is quite flexible in fitting the points and reducing RMSE, it distorts areas between the control points. This is shown by the poor match between image and vector roads (above right). Although it has a higher RMSE, the first-order transformation on the left is better overall.

points have been established using traditional surveying techniques. Indeed, the development of this "control network" infrastructure is one of the first and most important responsibilities of government. These survey points form the basis for distance, location, and area measurements used to define property, political, and municipal boundaries. As a result, this control network underlies most commerce, transportation, and land ownership and management. Coordinates, general location, and descriptions are documented for these control networks, and may be obtained from a number of government sources. In the United States, these sources include county surveyors, state surveyors, departments of transportation, and the National Geodetic Survey (NGS).

The ground survey network is often quite sparse and insufficient for registering many large-scale maps or images. Many may not be suitable for use as control points in a coordinate transformation of spatial data. First, our control points must be visible on the map, data layer, or image that we wish to register; and second, we must have precise ground coordinates in our target map projection. The first requirement, visibility on the source map or photograph, is often not met for survey-defined control. Therefore, we must often obtain additional control points.

Figure 4-28: Previous surveys are a common source of control points.

GNSS Control Points

The global positioning system (GPS), GLONASS, and Galileo are Global Navigation Satellite Systems (GNSS) that allow us to establish control points. GNSS, discussed in detail in Chapter 5, can help us obtain the coordinates of control points that are visible on a map or image. GNSS are particularly useful because we may quickly survey widely spaced points. GNSS positional accuracy depends on the technology and methods employed; it typically ranges from subcentimeter (tenths of inches) to a few meters (tens of feet). Most points recently added to the NGS and other government-maintained networks were measured using GNSS technologies.

Control Points from Existing Digital Data and Maps

Registered digital image data are common sources of ground control points, particularly when natural resources or municipal databases are to be developed for managing large areas. Digital images often provide a richly detailed depiction of surface features (Figure 4-29). Digital image data may be obtained that are registered to a known coordinate system. Typically, the coordinates of a corner pixel are provided, and the lines and columns for the image run parallel to the easting (E) and northing (N) direction of the coordinate system. Because the pixel dimensions are known, the calculation of a pixel coordinate involves multiplying the row and column number by the pixel size, and applying the corner offset, either by addition or subtraction. In this manner, the image row/column may be converted to an E, N coordinate pair, and control point coordinates determined.

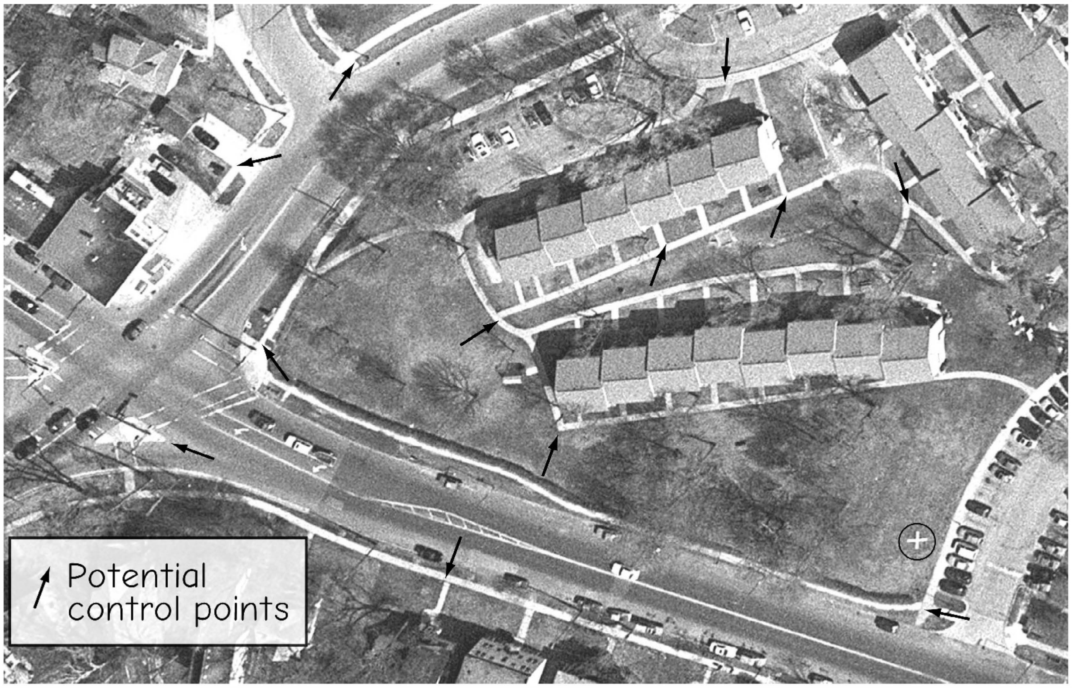


Figure 4-29: Potential control points, indicated here by arrows, may be extracted from digital images that have been geometrically corrected to provide an accurate, coordinate-projected rendering.

Existing maps are another common source of control points. Point locations are plotted and coordinates often printed on maps; for example the corner location coordinates are printed on USGS quadrangle maps. Road intersections and other well-defined locations are often represented on maps. If enough recognizable features can be identified, then control points may be obtained from the maps. Control points derived in this manner typically come only from cartometric maps, those maps produced with the intent of giving an accurate, map-projected representation of features on the Earth's surface.

Existing digital data may also provide control points. A short description of these digital data sources are provided here, and expanded descriptions of these and other digital data are provided in Chapter 7. For example, the USGS has produced Digital Raster Graphic (DRG) files that are scanned images of the 1:24,000-scale quadrangle maps. These DRGs come referenced to a standard coordinate system, so it is a simple and straightforward task to extract the coordinates of road intersections or other well-defined features that have been plotted on the USGS quadrangle maps. Vector data of roads are often widely available and, if of sufficient accuracy, may be used as a source of control points at road intersections and other distinct locations.

Raster Geometry and Resampling

Data often must be *resampled* when converting between coordinate systems, or changing the cell size of a raster data set (Figure 4-30). Resampling involves reassigning the cell values when changing raster coordinates or geometry. Resampling is required when changing cell sizes because the new cell centers will not align exactly with old cell centers. Changing coordinate systems may change the direction of the x and y axes, and GIS systems often require that the cell edges align with the coordinate system axes. Hence, the new cells often do not correspond to the same locations or extents as the old cells.

Common resampling approaches include the *nearest neighbor* (taking the output layer value from the nearest input layer cell center), *bilinear interpolation* (distance-based averaging of the four nearest cells), and *cubic convolution* (a weighted average of the sixteen nearest cells, Figure 4-30).

An example of a bilinear interpolation is shown in Figure 4-31. This algorithm uses a distance-weighted average of the four nearest cells in the input to calculate the value for the output. The new output location is represented by the black post. Initially, the height, or Z_{out} value, of the output location is unknown. Z_{out} is calculated based on the distances between the output locations and the input locations. The distance in the x direction is denoted in Figure 4-31 by d_1, and the distance in the y direction by d_2. The values in the input are shown as gray posts and are labeled as Z_1 through Z_4. Intermediate heights Z_b and Z_u are shown. These represent the average of the input values when taken in pairs in the x direction. These pairs are Z_1 and Z_2, to yield Z_u, and Z_3 and Z_4, to yield Z_b. Z_u and Z_b are then averaged to calculate Z_{out}, using the distance d_2 between the input and output locations to weight values at each input location. The cubic convolution resampling calculation is similar, except that more cells are used, and the weighting is not an average based on linear distance.

Map Projection vs. Transformation

Map transformations should not be confused with map projections. A map transformation typically employs a statistically fit linear equation to convert coordinates from one Cartesian coordinate system to another. A map projection, described in Chapter 3, differs from a transformation in that it is an analytical, formula-based conversion between coordinate systems, usu-

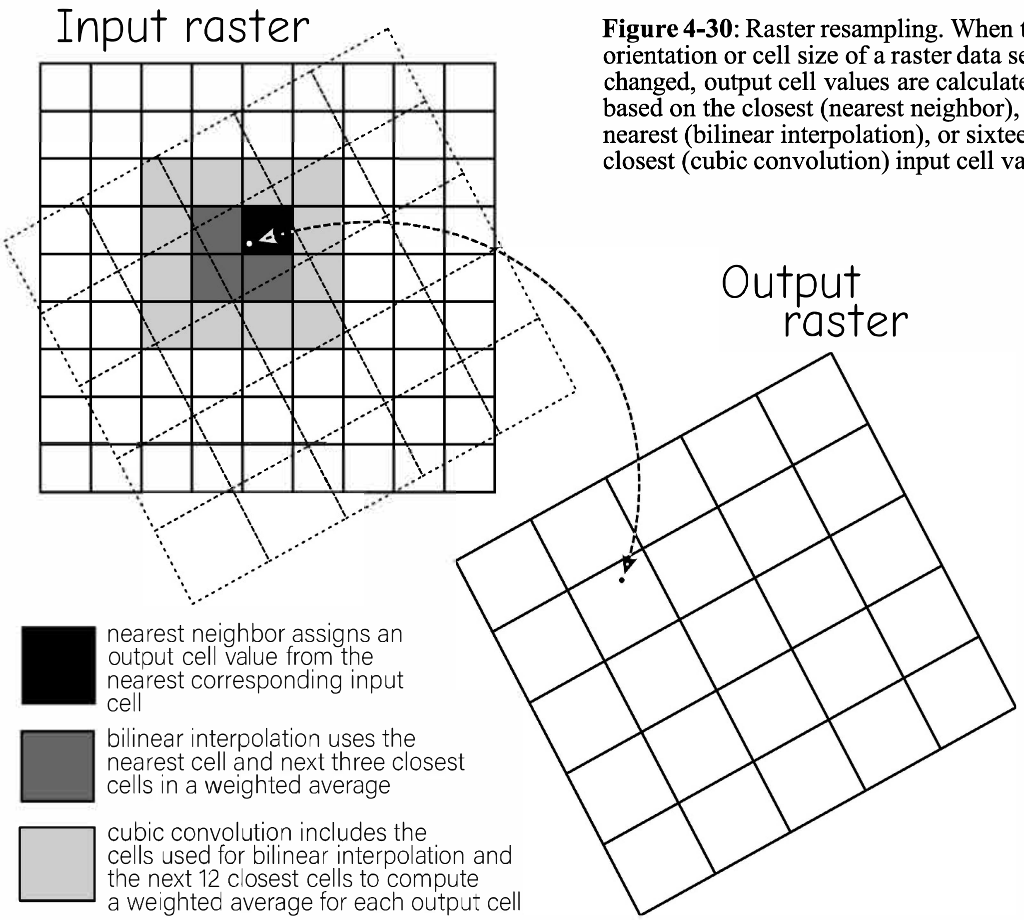
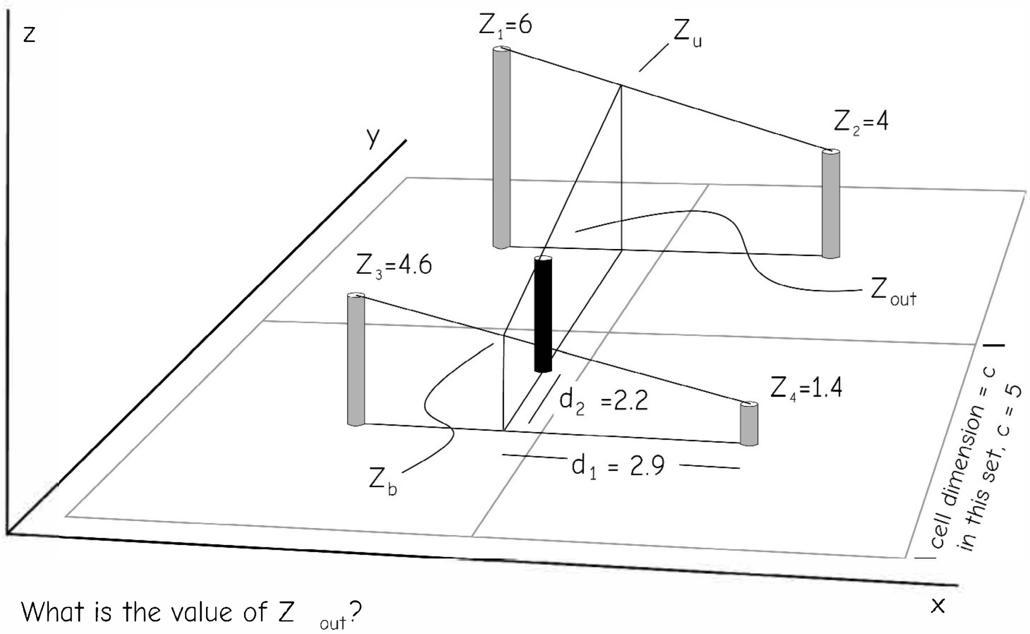


Figure 4-30: Raster resampling. When the orientation or cell size of a raster data set is changed, output cell values are calculated based on the closest (nearest neighbor), four nearest (bilinear interpolation), or sixteen closest (cubic convolution) input cell values.



What is the value of Z_{out}?

$$Z_b = Z_4 + \frac{(Z_3 - Z_4)*d_1}{c} \qquad Z_b = 1.4 + \frac{(4.6 - 1.4)*2.9}{5} = 3.26$$

$$Z_u = Z_2 + \frac{(Z_1 - Z_2)*d_1}{c} \qquad Z_u = 4 + \frac{(6 - 4)*2.9}{5} = 5.16$$

$$Z_{out} = Z_b + \frac{(Z_u - Z_b)*d_2}{c} \qquad Z_{out} = 3.26 + \frac{(5.16 - 3.26)*2.2}{5} = 4.1$$

Figure 4-31: The bilinear interpolation method uses a distance-weighted average to assign the output value, Zout, based on input values, Z_1 through Z_4.

ally from a curved, latitude/longitude coordinate system to a Cartesian coordinate system. No statistical fitting process is used with a map projection.

Map transformations should rarely be used in place of map projection equations when converting geographic data between map projections. Consider an example when data are delivered to an organization in Universal Transverse Mercator (UTM) coordinates and are to be converted to State Plane coordinates prior to integration into a GIS database. Two paths may be chosen. The first involves projection from UTM to geographic coordinates (latitude and longitude), and then from these geographic coordinates to the appropriate State Plane coordinates. This is the correct, most accurate approach.

An alternate and often less accurate approach involves using a transformation to convert between different map projections. In this case, a set of control points would be identified and the coordinates determined in both UTM and State Plane coordinate systems. The transformation coefficients would be estimated and these equations applied to all data in the UTM data layer. This new output data layer would be in State Plane coordinates. This transformation process should be avoided, as a transformation may introduce additional positional error.

Transforming between projections is used quite often, inadvertently, when digitizing data from paper maps. For example, USGS 1:24,000-scale maps are cast on a polyconic projection. If these maps are dig-

itized, it would be preferable to register them to the appropriate polyconic projection, and then reproject these data to the desired end projection. This is often not done, because the error in ignoring the projection over the size of the mapped area is typically less than the positional error associated with digitizing. Experience and specific calculations have shown that the spatial errors in using a transformation instead of a projection are small at these map scales under typical digitizing conditions.

This second approach, using a transformation when a projection is called for, should not be used until it has been tested as appropriate for each new set of conditions. Each map projection distorts the surface geometry. These distortions are complex and nonlinear. Affine or polynomial transformations are unlikely to remove this nonlinear distortion. Exceptions to this rule occur when the area being transformed is small, particularly when the projection distortion is small relative to the random uncertainties, transformation errors, or errors in the spatial data. However, there are no guidelines on what constitutes a sufficiently "small" area. In our example above, USGS 1:24,000 scale maps are often digitized directly into a UTM coordinate system with no obvious ill effects, because the errors in map production and digitizing are often much larger than those in the projection distortion for the map area. However, you should not infer this practice is appropriate under all conditions, particularly when working with smaller-scale maps.

Output: Hardcopy Maps, Digital Data, and Metadata

We create spatial data to use, share, and archive. Maps are often produced during data creation and distribution, as intermediate documents while editing, for analysis, or as finished products to communicate some aspect of our data. To be widely useful, we must also generate information, or "metadata," about the spatial data we've created, and we may have to convert our data to standard forms. This section describes some characteristics of data output. We start with a brief treatment of cartography and map design, by which we produce hardcopy and digital maps. We then provide a description of metadata, and some observations on data conversion and data transfer standards.

Cartography and Map Design

Cartography is the art and techniques of making maps. It encompasses both mapmaking tools and how these tools may be combined to communicate spatial information. Cartography is a discipline of much depth and breadth, and there are many books, journal articles, conferences, and societies devoted to the science and art of cartography. Our aim in the next few pages is to provide a brief overview of cartography with a particular focus on map design. This is both to acquaint new students with the most basic concepts of cartography, and help them apply these concepts in the consumption and production of spatial information. Readers interested in a more complete treatment should consult the references listed at the end of this chapter.

A primary purpose of cartography is to communicate spatial information. This requires identification of the

-intended audience,

-information to communicate,

-area of interest,

-physical and resource limitations;

in short, the whom, what, where, and how we may present our information.

These considerations drive the major cartographic design decisions we make each time we produce a map. We must consider the:

-scale, size, shape, and other general map properties,

-data to plot,

-symbol shapes, sizes, or patterns,

-labeling, including font type and size,

-legend properties, size, and borders, and

-the placement of all these elements on a map.

Map scale, size, and shape depend primarily on the intended map use. Wall maps for viewing at a distance of a meter or more may have few, large, boldly colored features. In contrast, commonly produced street maps for navigation in metropolitan areas are detailed, to be viewed at short ranges, and have a rich set of additional tables, lists, or other features.

Map scale is often determined in part by the size of the primary objects we wish to display, and in part by the most appropriate media sizes, such as the page or screen size possible for a document. As noted earlier, the map scale is the ratio of lengths on a map to true lengths. If we wish to display an area that spans 25 km (25,000 m) on a screen that spans 25 cm (0.25 m), the map scale will be near 0.25 to 25,000, or 1:100,000. This decision on size, area, and scale then drives further map design. For example, scale limits the features we may display, and the size, number, and labeling of features. At a 1:100,000 scale we may not be able to show all cities, burgs, and towns, as there may be too many to fit at a readable size.

Maps typically have a primary theme or purpose that is determined by the intended audience. Is the map for a general population, or for a target audience with specific expectations for map features and design? General purpose maps typically have a wide

range of features represented, including transportation networks, towns, elevation, or other common features (Figure 4-32a). Special purpose maps, such as road maps, focus on a more limited set of features, in this instance road locations and names, town names, and large geographic features (Figure 4-32b).

Once the features to include on a map are defined, we must choose the symbols used to draw them. Symbology depends in part on the type of feature. For example, we have a different set of options when representing continuous features such as elevation or pollution concentration than when representing discrete features. We also must choose among symbols for each of the types of discrete features; for example, the set of symbols for points are generally different from those for line or area features.

Symbol size is an important attribute of map symbology, often specified in a unit called a *point*. One point is approximately equal to 0.467 mm, or about 1/55th of an inch. A specific point number is most often used to specify the size of symbols, for example, the dimensions of small squares to represent houses on a map, or the characteristics of a specific pattern used to fill areas on a map. A line width may also be specified in points. Setting a line width of two points means we want that particular line plotted with a width of 0.93 mm. It is unfortunate that "point" is both the name of the distance unit and a general property of a geographic feature, as in "a tree is a point feature." This forces us to talk about the "point size" of symbols to represent points, lines, or area fills or patterns, but if we are careful, we may communicate these specifications clearly.

The best size, pattern, shape, and color used to symbolize each feature depends on the viewing distance; the number, density, and type of features; and the purpose of the map. Generally, we use larger, bolder, or

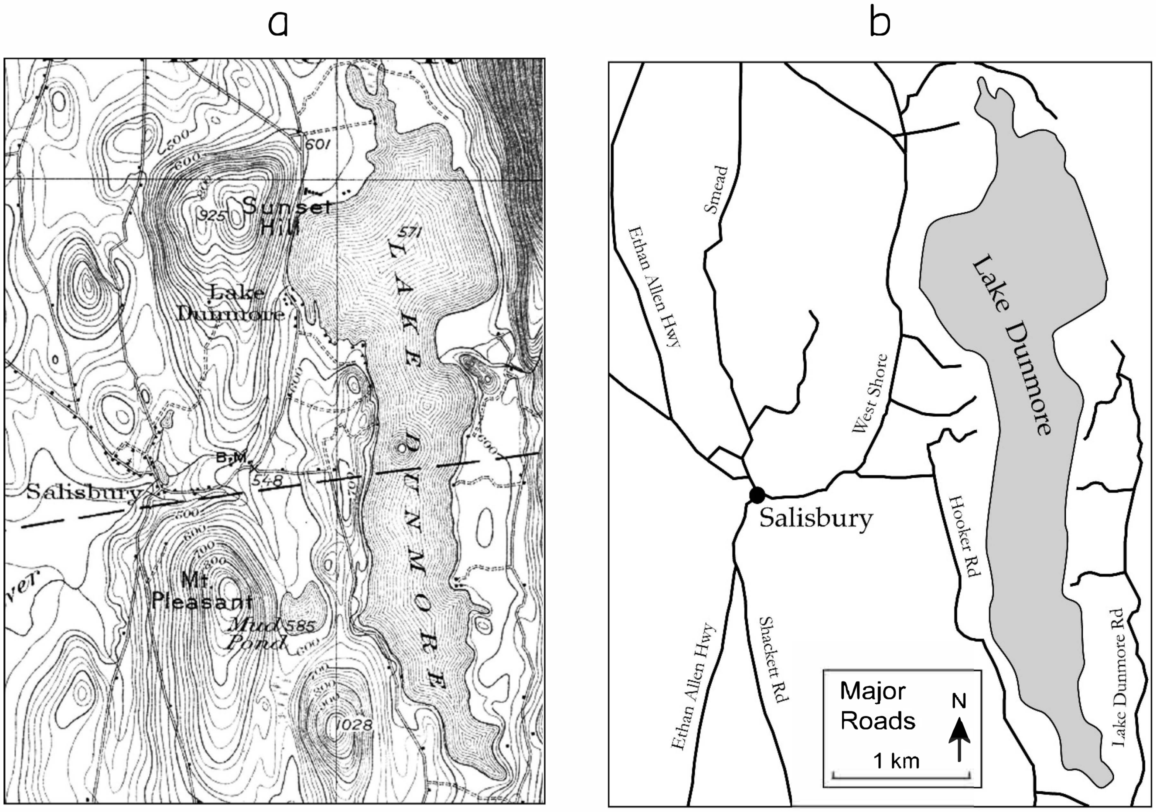


Figure 4-32: Example of a) a detailed, general-purpose map, here a portion of a United States Geological Survey map, and b) a specialized map focusing on a specific set of selected features, here showing roads. The features chosen for depiction on the map depend on the intended map use.

thicker symbols for maps to be viewed from longer distances, while we reduce this limit when producing maps for viewing at 50 cm (20 in). Most people with normal vision under good lighting may resolve lines down to near 0.2 points at close distances, provided the lines show good contrast with the background. Although size limits depend largely on background color and contrast, point features are typically not resolvable at sizes smaller than about 0.5 points, and distinguishing between shapes is difficult for point features smaller than approximately 2 points in their largest dimension.

The pattern and color of symbols must also be chosen, generally from a set provided by the software (Figure 4-33). Symbols generally distinguish among feature type by characteristics, and although most symbols are not associated with a feature type, some are, such as, plane outlines for airports, numbered shields for highways, or a hatched line for a railroad.

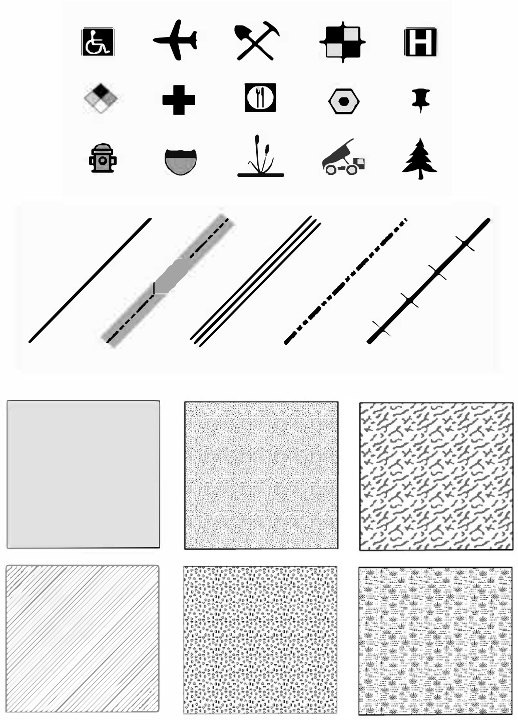


Figure 4-33: Examples of point (top), line (mid), and area (bottom) symbols used to distinguish among features of different types. Most GIS software provides a set of standard symbols for point, line, area, and continuous surface features.

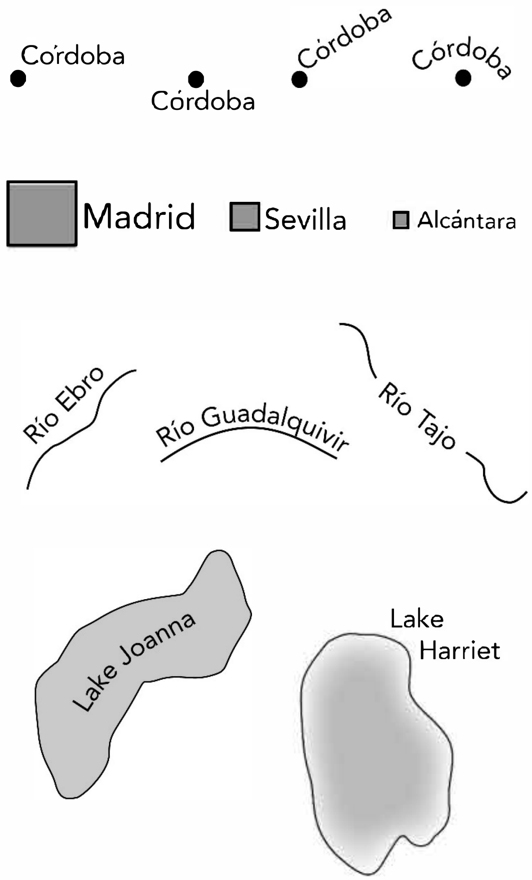


Figure 4-34: Common labeling options, including straight, angled, wrapped text, and graduated labels for points (top two sets), and angled, wrapped, fronting, and embedded labels for line and polygon features (bottom two sets).

We also must often choose whether and how to label features. Most GIS software provides a range of tools for creating and placing labels, and in all cases we must choose the label font type and size, location relative to the feature, and orientation. Primary considerations when labeling point features are label placement relative to the point location, label size, and label orientation (Figure 4-34). We may also use graduated labels, that is, resize them according to some variable associated with the point feature. For example, it is common to have larger features and label fonts for larger cities (Figure 4-34). Labels may be bent, angled, or wrapped around features to

improve clarity and more efficiently use space in a map.

Label placement is very much an art, and there is often much individual editing required when placing and sizing labels for finished maps. Most software provides for automatic label placement, usually specified relative to feature location. For example, one may specify labels above and to the right of all points, or line labels placed over line features, or polygon labels placed near the polygon centroid. However, these automatic placements may not be satisfactory because labels may overlap, labels may fall in cluttered areas of the map, or features associated with labels may be ambiguous. Some software provides options for automatic label placement, including removal or movement of overlapping labels. These often reduce manual editing, but sometimes increase it.

Figure 4-35 shows a portion of a map of southern Finland. This region presents several mapping problems, including the high density of cities near the upper right, an irregular coastline, and dense clustering of islands along the coast. Most labels are placed above and to the right of their corresponding city; however, some are moved or angled for clarity. Cities near the coast show both, to avoid labels crossing the water/land boundary where practical. Semitransparent background shading is added for Parainen and Hanko, cities placed in the island matrix. This example demonstrates the individual editing often required when placing labels.

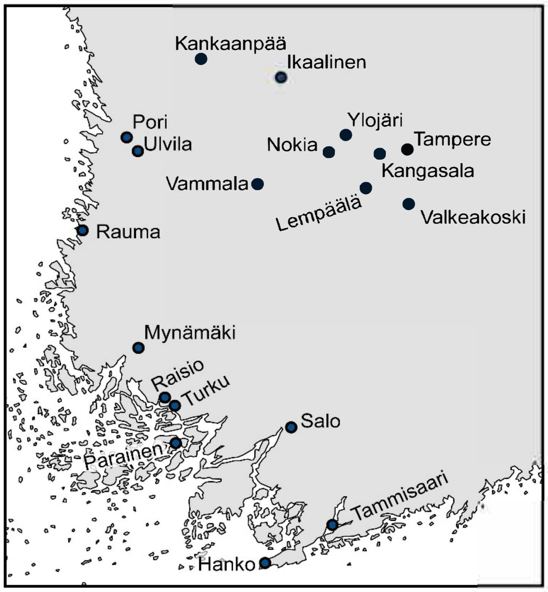


Figure 4-35: Example label placement for cities in southern Finland.

Most maps should have legends. The legend identifies map features succinctly and describes the symbols used to depict those features. Legends often include or are grouped with additional map information such as scale bars, north arrows, and descriptive text. The cartographer must choose the size and shape of the descriptive symbol, and the font type, size, and orientation for each symbol in the legend. The primary goal is to have a clear, concise, and complete legend.

The kind of symbols appropriate for map legends depends on the types of features depicted. Different choices are available for point, line, and polygon features, or for continuously variable features stored as rasters. Most software provides a range of legend elements and symbols that may be used. Typically, these tools allow a wide range of symbolizations, and a compact way of describing the symbolization in a legend (Figure 4-36).

The specific layout of legend features must be defined; for example, the point feature symbol size may be graduated based on some attribute for the points. Successively larger features may be assigned for successively larger cities. This must be noted in the legend, and the symbols nested, shown sequentially, or otherwise depicted (Figure 4-36, top left).

The legend should be exhaustive. Examples of each different symbol type that appears on the map should appear in the legend. This means each point, line, or area symbol is drawn in the legend with some descriptive label. Labels may be next to, wrapped around, or embedded within the features, and sometimes descriptive numbers

are added, for example, a range of continuous variables (Figure 4-36, upper left). Scale bars, north arrows, and descriptive text boxes are typically included in the legend.

Map composition or layout is another primary task. Composition consists of determining the map elements, their size, and their placement. Typical map elements, shown in Figure 4-3 and Figure 4-4, include one or more main data panes or areas, a legend, a title, a scale bar and north arrow, a grid or graticule, and perhaps descriptive text. These each must be sized and placed on the map.

These map elements should be positioned and sized in accordance with their importance. The map's most important data pane should be largest, and it is often centered or otherwise given visual dominance. Other elements are typically smaller and located around the periphery or embedded within the main data pane. These other elements include map insets, which are smaller data panes that show larger or smaller scale views of a region in the primary data pane. Good map compositions usually group related elements and use empty space effectively. Data panes are often grouped and legend elements placed near each other, and grouping is often indicated with enclosing boxes.

Neophyte cartographers should avoid two tendencies in map composition, both depicted in Figure 4-37. First, it is generally easy to create a map with automatic label and legend generation and placement. The map shown at the top of Figure 4-37 is typical of this automatic composition, and includes poorly placed legend elements and too small, poorly placed labels. Labels crowd each other, are ambiguous, and cross water/land or other feature boundaries, and fonts are poorly chosen. You should note that automatic map symbol selection and placement is nearly always suboptimal, and the novice cartographer should scrutinize these choices and manually improve them.

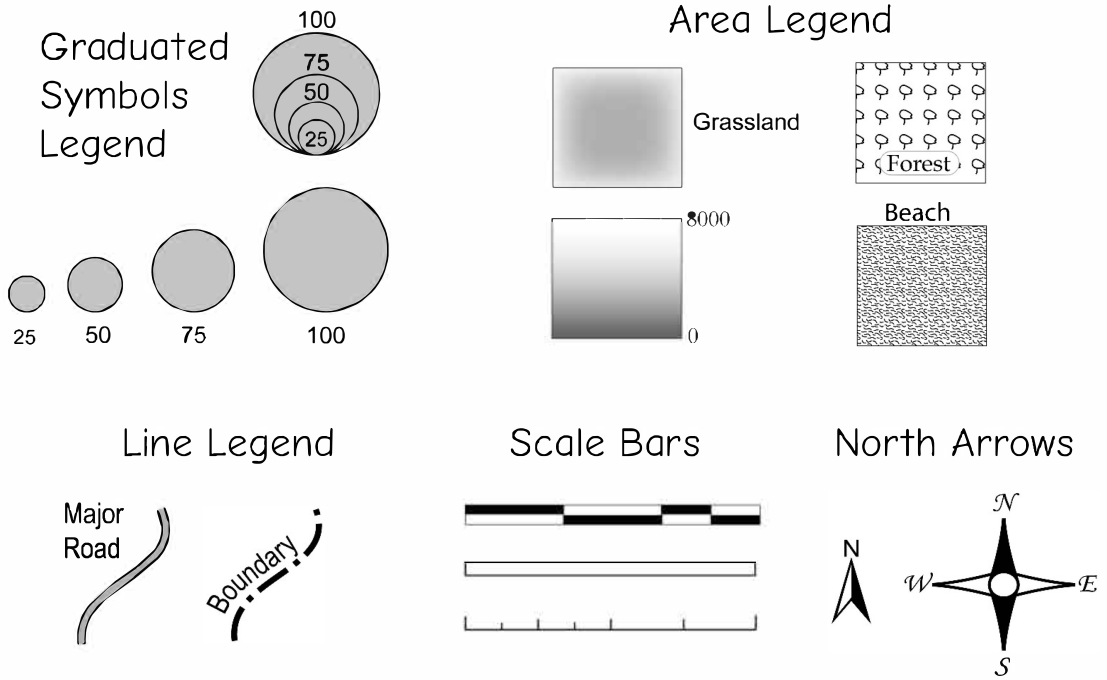


Figure 4-36: Examples of legend elements and representation of symbols. Some symbols may be grouped in a compact way to communicate the values associated with each symbol, e.g., sequential or nested graduated circles to represent city population size, area pattern or color fills to distinguish among different polygon features, line and point symbols, and informative elements such as scale bars and north arrows.

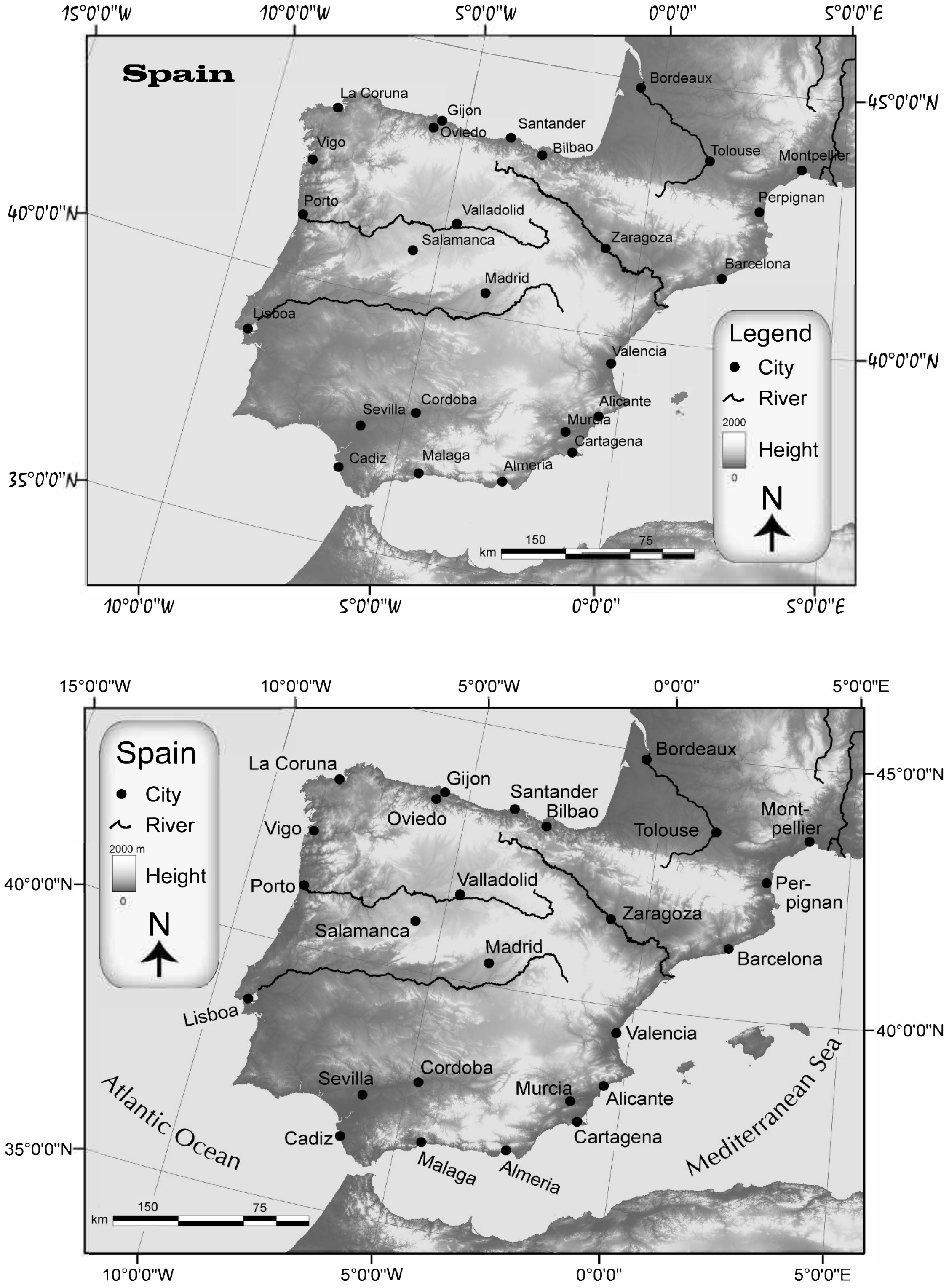


Figure 4-37: An example of poor map design (top). This top panel shows a number of mistakes common for the neophyte cartographer, including small labels (cities) and mismatched fonts (graticule labels, title), poor labeling (city labels overlapping, ambiguously placed, and crossing distinctly shaded areas), unlabeled features (oceans and seas), poorly placed scale bar and legend, and unbalanced open space on the left side of the map. These problems are not present in the improved map design, shown in the lower panel.

The second common error is poor use of empty space, those parts of the map without map elements. There are two opposite tendencies: either to leave too much or unbalanced empty space or to clutter the map in an attempt to fill all empty space. Note that the map shown at the top of Figure 4-37 leaves large empty spaces on the left (western) edge, with the Atlantic Ocean devoid of features. The cartographer may address this in several ways: by changing the size, shape, or extent of the area mapped; by adding new features, such as data panes as insets, additional text boxes, or other elements; or by moving the legend or other map elements to that space. The map shown at the bottom of Figure 4-37, while not perfect, fixes these design flaws, in part by moving the legend and scale bar, and in part by adding labels for the Atlantic Ocean and Mediterranean Sea. The empty space is more balanced in that it appears around the major map elements in approximately equal proportions.

As noted earlier, this is only a brief introduction to cartography, a subject covered by many good books, some listed at the end of this chapter. Perhaps the best compendium of examples is the Map Book Series, by ESRI, published annually since 1984. Examples are available at the time of this writing at www.esri.com/mapmuseum. You should leaf through several volumes in this series, with an eye toward critical map design. Each volume contains many beautiful and informative maps, and provides techniques worth emulating.

Digital Data Output

We often must transfer digital data we create to another user, or use data developed by others. Given the number of different GIS software, operating systems, and computer types, transferring data is not always a straightforward process. Digital data output typically includes two components, the data themselves in some standard, defined format, and *metadata*, or data about the digital data. We will describe data formats and metadata in turn.

Digital data are the data in some electronic form. As described at the end of the first chapter, there are many file formats, or ways of encoding the spatial and attribute data in digital files. Digital data output often consists of recording or converting data into one of these file formats. These data are typically converted with a utility, tool, or option available in the data development software (Figure 4-38). The most useful of these utilities supports a broad range of input and output options, each fully described in the program documentation.

All formats strive for complete data transfer without loss. They must transmit the spatial and attribute data, the metadata, and all other information necessary to effectively use the spatial data. There are many digital data output formats, although many are legacy formats that are used with decreasing frequency.

A common contemporary format is the *Geographic Markup Language (GML)*. This is an extension of XML for geographic features; XML is in turn the lingua franca for human/machine readable documents. As with most XML, there are two parts for any

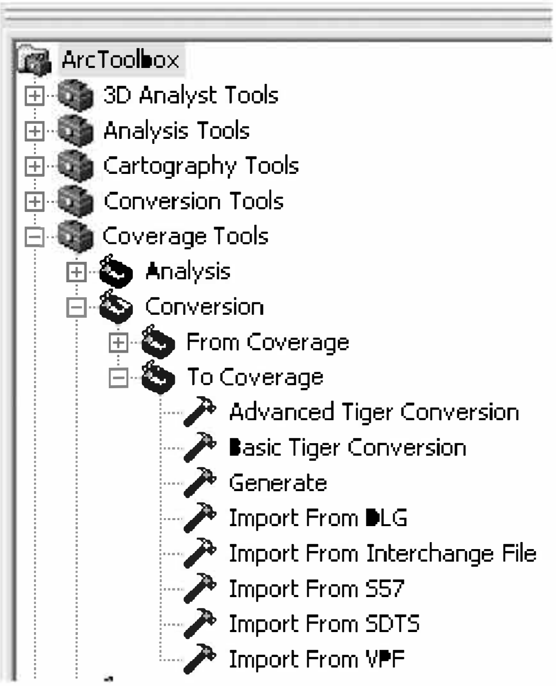


Figure 4-38: An example of a conversion utility, here from the ESRI ArcGIS software. Date may be converted from one of several formats to an ESRI-specific digital data.

GML dataset: a schema that describes the document, and the document containing the geographic data. GML is a standard, but there can be many extensions, so a community of users can extend the standard with additional features, and document the extension in a standard way.

There are many legacy digital data transfer formats that were widely used before GML. GML replaced a withdrawn standard, the *Spatial Data Transfer Standard* (SDTS), with translators available to or from this older format. There are several U.S. Geological Survey formats for the transfer of digital elevation models or digital vector data, or software-specific formats, such as an ASCII format (GEN/UNGEN) that was developed by ESRI. These were useful for a limited set of transfers, but shortcomings in each of these transfer formats led to the development of subsequent standards. These formats are not common, but sometimes arise in converting older data sets.

Metadata: Data Documentation

Metadata are information about spatial data. Metadata describe the content, source, lineage, methods, developer, coordinate system, extent, structure, spatial accuracy, attributes, and responsible organization for spatial data.

Metadata are required for the effective use of spatial data. Metadata allow the efficient transfer of information about data, and inform new users about the geographic extent, coordinate system, quality, and other data characteristics. Metadata aid organizations in evaluating data to determine if they are suitable for an intended use, e.g., to review accuracy, coverage, or information needs. Metadata may also aid in data updates by guiding the choice of appropriate collection methods and formats for new data.

Most governments have or are in the process of establishing standard methods for reporting metadata. In the United States, the Federal Geographic Data Committee (FGDC) has defined a Content Standard for Digital Geospatial Metadata (CSDGM) to specify the content and format for metadata. The CSDGM ensures that spatial data are clearly described so that they may be used effectively within an organization. The use of the CSDGM also ensures that data may be described to other organizations in a standard manner, and that spatial data may be more easily evaluated by and transferred to other organizations.

The CSDGM consists of a standard set of elements that are presented in a specified order. The standard is exhaustive in the information it provides, and is flexible in that it may be extended to include new elements for new categories of information in the future. There are over 330 different elements in the CSDGM. Some of these elements contain information about the spatial data, and some elements describe or provide linkages to other elements. Elements have standardized long and short names and are provided in a standard order with a hierarchical numbering system. For example, the westernmost bounding coordinate of a data set is element 1.5.1.1, defined as follows:

1.5.1.1 West Bounding Coordinate – westernmost coordinate of the limit of coverage expressed in longitude.

Type: real

Domain: -180.0 < = West Bounding Coordinate < 180.0

Short Name: westbc

The numbering system is hierarchical. Here, 1 indicates it is basic identification information, 1.5 indicates identification information about the spatial domain, 1.5.1 is for bounding coordinates, and 1.5.1.1 is the western most bounding coordinate.

There are 10 basic types of information in the CSDGM:

1) identification, describing the data set,
2) data quality,
3) spatial data organization,
4) spatial reference coordinate system,
5) entity and attribute,

6) distribution and options for obtaining the data set,

7) currency of metadata and responsible party,

8) citation,

9) time period information, used with other sections to provide temporal information, and

10) contact organization or person.

The CSDGM is a content standard and does not specify the format of the metadata. As long as the elements are included, properly numbered, and identified with correct values describing the data set, the metadata are considered to conform with the CSDGM. Indentation and spacing are not specified. However, because metadata may be quite complex, there are a number of conventions that are emerging in the presentation of metadata. These conventions seek to ensure that metadata are presented in a clear, logical way to humans, and are also easily ingested by computer software. There is a Standard Generalized Markup Language (SGML) for the exchange of metadata. An example of a portion of the metadata for a 1:100,000 scale digital line graph data set is shown in Figure 4-39.

Metadata are most often created using specialized software tools. Although meta-

```
4. Spatial_Reference_Information:
  4.1 Horizontal_Coordinate_System_Definition:
    4.1.2 Planar:
      4.1.2.2 Grid_Coordinate_System:
        4.1.2.2.1 Grid_Coordinate_System_Name:
                          Universal Transverse Mercator
        4.1.2.2.2 Universal_Transverse_Mercator:
          4.1.2.2.2.1 UTM_Zone_Number: 10-19
      4.1.2.4 Planar_Coordinate_Information:
        4.1.2.4.1 Planar_Coordinate_Encoding_Method:
                                   coordinate pair
        4.1.2.4.2 Coordinate_Representation:
          4.1.2.4.2.1 Abscissa_Resolution: 2.54
          4.1.2.4.2.2 Ordinate_Resolution: 2.54
        4.1.2.4.4 Planar_Distance_Units: meters
    4.1.4 Geodetic_Model:
      4.1.4.1 Horizontal_Datum_Name: North American Datum 1927
      4.1.4.2 Ellipsoid_Name: Clark 1866
      4.1.4.3 Semi-major_Axis: 6378206.4
      4.1.4.4 Denominator_of_Flattening_Ratio: 294.98
  4.2 Vertical_Coordinate_System_Definition:
    4.2.1 Altitude_System_Definition:
      4.2.1.1 Altitude_Datum_Name:
                   National Geodetic Vertical Datum of 1929
      4.2.1.2 Altitude_Resolution: 1
      4.2.1.3 Altitude_Distance_Units: feet or meters
      4.2.1.4 Altitude_Encoding_Method: attribute values
    4.2.2 Depth_System_Definition:
      4.2.2.1 Depth_Datum_Name: Mean lower low water
      4.2.2.2 Depth_Resolution: 1
      4.2.2.3 Depth_Distance_Units: meters or feet
      4.2.2.4 Depth_Encoding_Method: attribute values
```

Figure 4-39: Example of a small portion of the FGDC recommended metadata for a 1:100,000 scale derived digital data set.

data may be produced using a text editor, the numbering system, names, and other conventions are laborious to type. There are often complex linkages between metadata elements, and some elements are repeated or redundant. Software tools may ease the task of metadata entry by reducing redundant entries, ensuring correct linkages, and checking elements for contradictory information or errors. For example, the metadata entry tool may check to make sure the westernmost boundary is west of the easternmost boundary. Metadata are most easily and effectively produced when their development is integrated into the workflow of data production.

Although not all organizations in the United States adhere to the CSDGM metadata standard, most organizations record and organize a description and other important information about their data, and many organizations consider a data set incomplete if it lacks metadata. All U.S. government units are required to adhere to the CSDGM when documenting and distributing spatial data.

Many other national governments are developing metadata standards. One example is the spatial metadata standard developed by the Australia and New Zealand Land Information Council (ANZLIC), known as the ANZLIC Metadata Guidelines. ANZLIC is a group of government, business, and academic representatives working to develop spatial data standards. The ANZLIC metadata guidelines define the core elements of metadata, and describe how to write, store, and disseminate these core elements. Data entry tools, examples, and spatial data directory have been developed to assist in the use of ANZLIC spatial metadata guidelines.

There is a parallel effort to develop and maintain international standards for metadata. The standards are known as the ISO 19115 International Standards for Metadata. According to the International Standards Organization, the ISO 19115 "defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data."

There is a need to reconcile international and national metadata standards, because they may differ. National standards may require information not contained in international standards, or vice versa. Governments typically create *metadata profiles* that are consistent with the international standard. These profiles establish the correspondence between elements in the different standards, and identify elements of the international profile that are not in the national profile.

Summary

Spatial data entry is a common activity for many GIS users. Although data may be derived from several sources, maps are a common source, and care must be taken to choose appropriate map types and to interpret the maps correctly when converting them to spatial data in a GIS.

Maps are used for spatial data entry due to several unique characteristics. These include our long history of hardcopy map production, so centuries of spatial information are stored there. In addition, maps are inexpensive, widely available, and easy to convert to digital forms, although the process is often time consuming, and may be costly. Maps are usually converted to digital data through a manual digitization process, whereby a human analyst traces and records the location of important features. Maps may also be digitized via a scanning device.

The quality of data derived from a map depends on the type and size of the map, how the map was produced, the map scale, and the methods used for digitizing. Large-scale maps generally provide more accurate positional data than comparable small-scale maps. Large-scale maps often have less map generalization, and small horizontal errors in plotting, printing, and digitizing are magnified less during conversion of large-scale maps.

Snapping, smoothing, vertex thinning, and other tools may be used to improve the quality and utility of digitized data. These methods are used to ensure positional data are captured efficiently and at the proper level of detail.

Map and other data often need to be converted to a target coordinate system via a map transformation. Transformations are different from map projections (discussed in Chapter 3), in that a transformation uses an empirical, least squares process to convert coordinates from one Cartesian system to another. Transformations are often used when registering digitized data to a known coordinate system. Map transformations should not be used when a map projection is called for.

Cartography is an important aspect of GIS, because we often communicate spatial information through maps. Map design depends on both the target audience and purpose, setting and modes of map viewing, and available resources. Proper map design considers the scale, symbols, labels, legend, and placement to effectively communicate the desired information.

Metadata are the "data about data." They describe the content, origin, form, coordinate system, spatial and attribute data characteristics, and other relevant information about spatial data. Metadata facilitate the proper use, maintenance, and transfer of spatial data. Metadata standards have been developed, both nationally and internationally, with profiles used to cross-reference elements between metadata standards. Metadata are a key component of spatial data, and many organizations do not consider data complete until metadata have been created.

Suggested Reading

Aronoff, S. (1989). *Geographic Information Systems, A Management Perspective*. Ottawa: WDL Publications.

Bolstad, P., Gessler, P., Lillesand, T.M. (1990). Positional uncertainty in manually digitized map data. *International Journal of Geographical Information Systems*, 4:399–412.

Burrough, P.A., Frank, A.U. (1996). *Geographical Objects with Indeterminate Boundaries*. London: Taylor & Francis.

Chrisman, N.R. (1984). The role of quality information in the long-term functioning of a geographic information system. *Cartographica*, 21:79-87.

Chrisman, N.R. (1987). Efficient digitizing through the combination of appropriate hardware and software for error detection and editing. *International Journal of Geographical Information Systems*, 1:265–277.

DeMers, M. (2000). *Fundamentals of Geographic Information Systems* (2nd ed.). New York: Wiley.

Douglas, D.H., Peuker, T.K. (1973). Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Canadian Cartographer*, 10:112–122.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., Tyler C. (2002). The National Elevation Dataset. *Photogrammetric Engineering and Remote Sensing*, 68:5–32.

Holroyd, F., Bell, S.B.M. (1992). Raster GIS: Models of raster encoding. *Computers and Geosciences*, 18:419–426.

Joao, E.M. (1998). *Causes and Consequences of Map Generalization.* London: Taylor & Francis.

Laurini, R., Thompson, D. (1992). *Fundamentals of Spatial Information Systems*. London: Academic Press.

Maquire, D.J., Goodchild, M.F., Rhind, D. (Eds.). (1991). *Geographical Information Systems: Principles and Applications*. Harlow: Longman Scientific.

McBratney, A.B., Santos, M.L.M., Minasny, B. (2003). On digital soil mapping. *Geoderma,* 117:3–52.

Muehrcke, P.C., Muehrcke, J.P. (1992). *Map Use: Reading, Analysis, and Interpretation* (3rd ed.). Madison: J.P. Publications.

Nagy, G., Wagle, S.G. (1979). Approximation of polygonal maps by cellular maps. *Communications of the Association of Computational Machinery*, 22:518–525.

Peuker, T.K., Chrisman, N. (1975). Cartographic data structures. *The American Cartographer*, 2:55–69.

Peuquet, D.J. (1984). A conceptual framework and comparison of spatial data models. *Cartographica*, 21:66–113.

Peuquet, D.J. (1981). An examination of techniques for reformatting digital cartographic data. Part II: the raster to vector process. *Cartographica*, 18:21–33.

Shaeffer, C.A., Samet, H., Nelson, R.C. (1990). QUILT: a geographic information system based on quadtrees. *International Journal of Geographical Information Systems*, 4:103–132.

Shea, K.S., McMaster, R.B. (1989). Cartographic generalization in a digital environment: when and how to generalize. *Proceedings AutoCarto 9*, 56–67.

Warner, W., Carson, W. (1991). Errors associated with a standard digitizing tablet. *ITC Journal*, 2:82–85.

Weibel, R. (1997). Generalization of spatial data: principles and selected algorithms. In van Kreveld, M., Nievergelt, J., Roos, T., Widmayer, P., (Eds.). *Algorithmic Foundations of Geographic Information Systems*. Berlin: Springer-Verlag.

Wolf, P.R., Ghilani, C. (2002). *Elementary Surveying, an Introduction to Geomatics* (10th ed.). New Jersey: Prentice-Hall.

Zeiler, M. (1999). *Modeling Our World: The ESRI Guide to Geodatabase Design*. Redlands: ESRI Press.

Study Questions

4.1 - Which is the larger-scale map:

a) 1:5,000, or 1:15,000?

b) 1:5,286 or 1 inch to a mile?

c) 1:1,000,000, or 1 cm to 1 km?

e) 1:50,000, or 0.00025

e) 5:1, or 1:1?

4.2 - Which is a larger-scale map,

a)1:20,000 or 1:1,000,000?

b) 1 centimeter to 1,000 meters or one yard to a mile

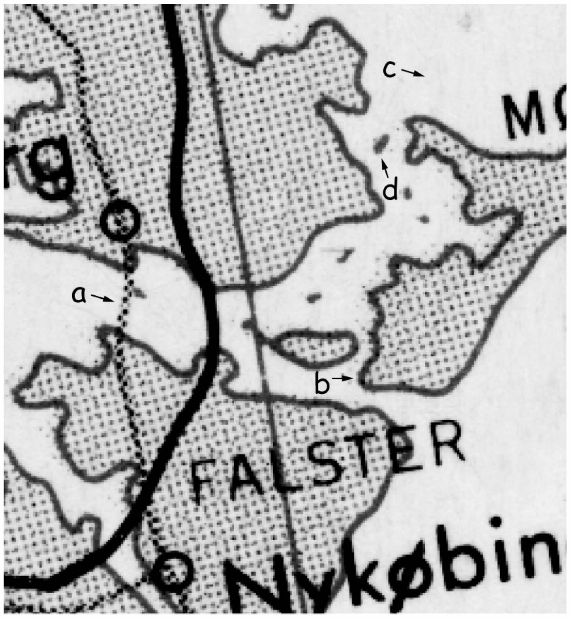
c) 1 inch equals 1 mile, or 1:100,000

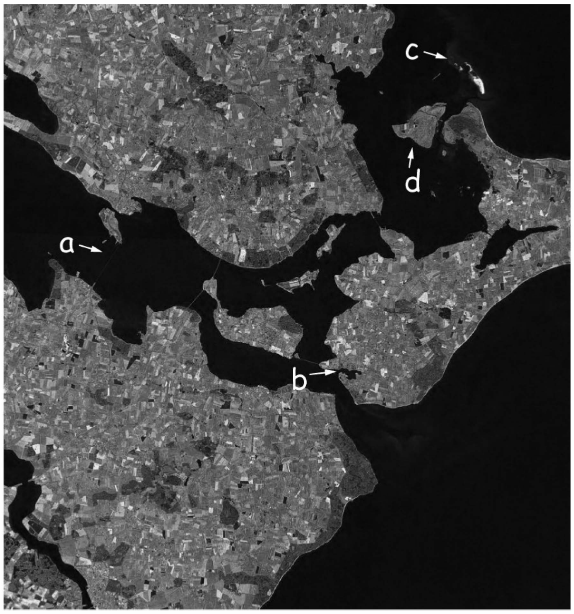
d) 1 cm to 1 km or 1 inch to a mile

e) 1 mm to 1 km, or f) 1:1,500,000

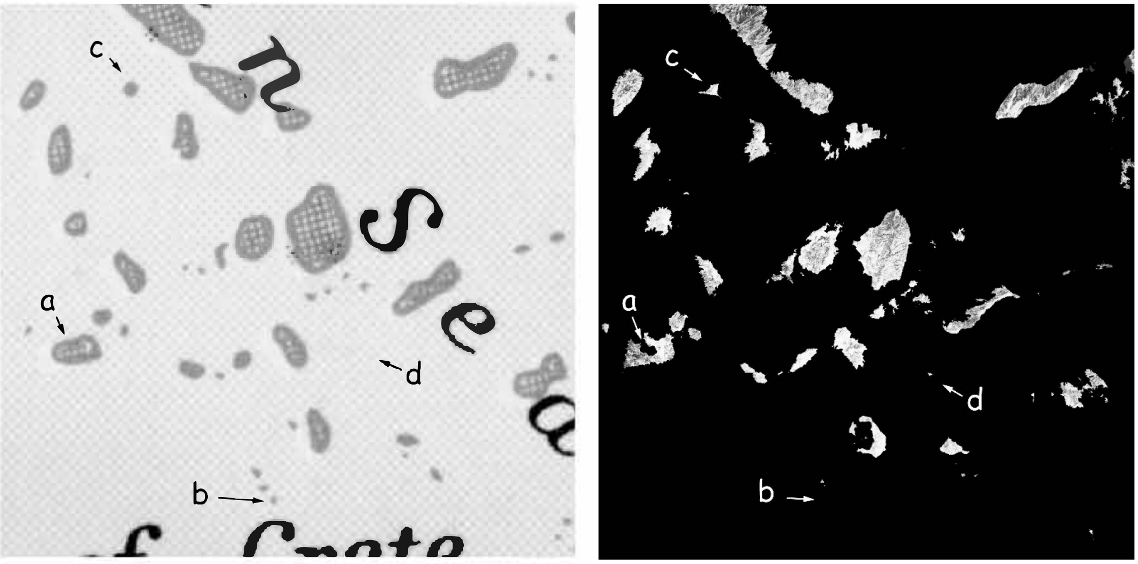
4.3 - Describe three different types of generalization.

4.4 - Identify the kind of generalization at the labeled locations a through d in the map below, left, compared to the "truth" in the image, below right. Categorize the generalizations as fused, simplified, displaced, omitted, or exaggerated.





4.5 - Identify the kind of generalization at the labeled locations a through d in the map below, left, compared to the "truth" in the image, below right. Categorize the generalizations as fused, simplified, displaced, omitted, or exaggerated; or if it doesn't fit in one of these categories, then categorize it as "other," and describe the generalization.



4.6 - What are the most common map media? Why?

4.7 - Is media deformation more problematic with large-scale maps or small scale paper maps? Why?

4.8 - Which map typically shows more detail – a large-scale map or a small-scale map? Can you give three reasons why?

4.9 - Complete the following table that shows scale measurements and calculations.

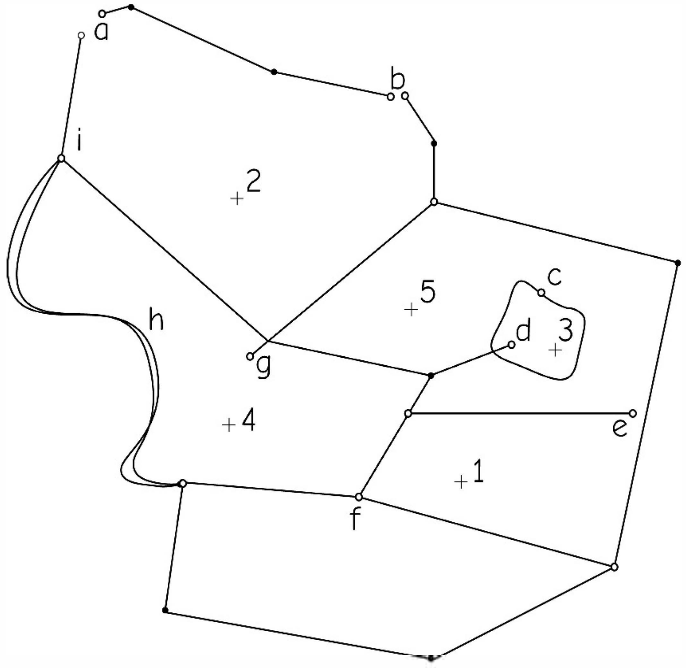
Ground distance and units	Corresponding map distance and units	Map Scale
13,280 feet	6.4 inches	1 : 24,900
126.4 kilometers	25.28 centimeters	
123.6 miles	22.8 inches	
40.7 meters	centimeters	1 : 502.5
kilometers	4.62 inches	1 : 249,685

4.10 - Complete the following table that shows scale measurements and calculations.

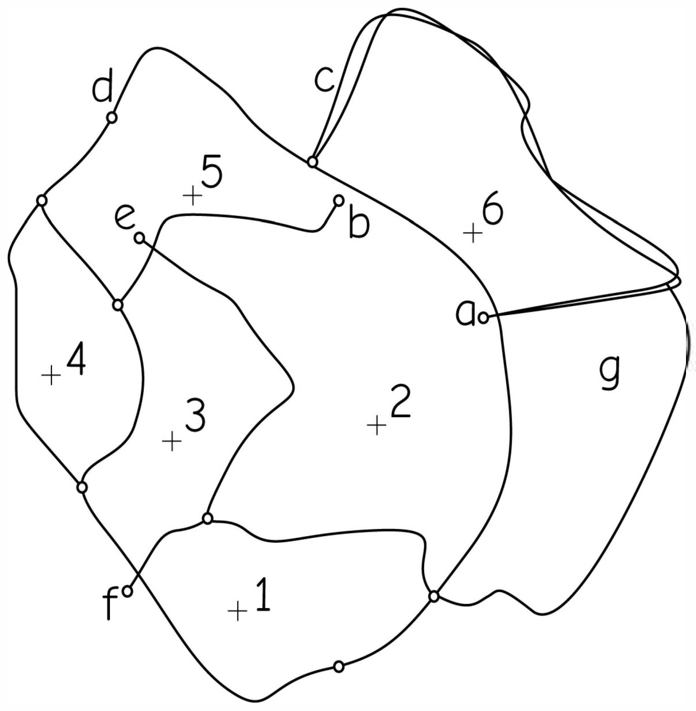
Ground distance and units	Corresponding map distance and units	Map Scale
17,120 kilometers	16.85 inches	1 : 40,000,935
23.4 kilometers	11.7 centimeters	1 : 200,000
16.4 miles	9.3 inches	1 : 109,869
102.0 meters	1.855 centimeters	1 : 5,500
320.08 miles	10.24 inches	1 : 2,000,000

4.11 - What is snapping in the context of digitizing? What are undershoots and overshoots, and why are they undesirable?

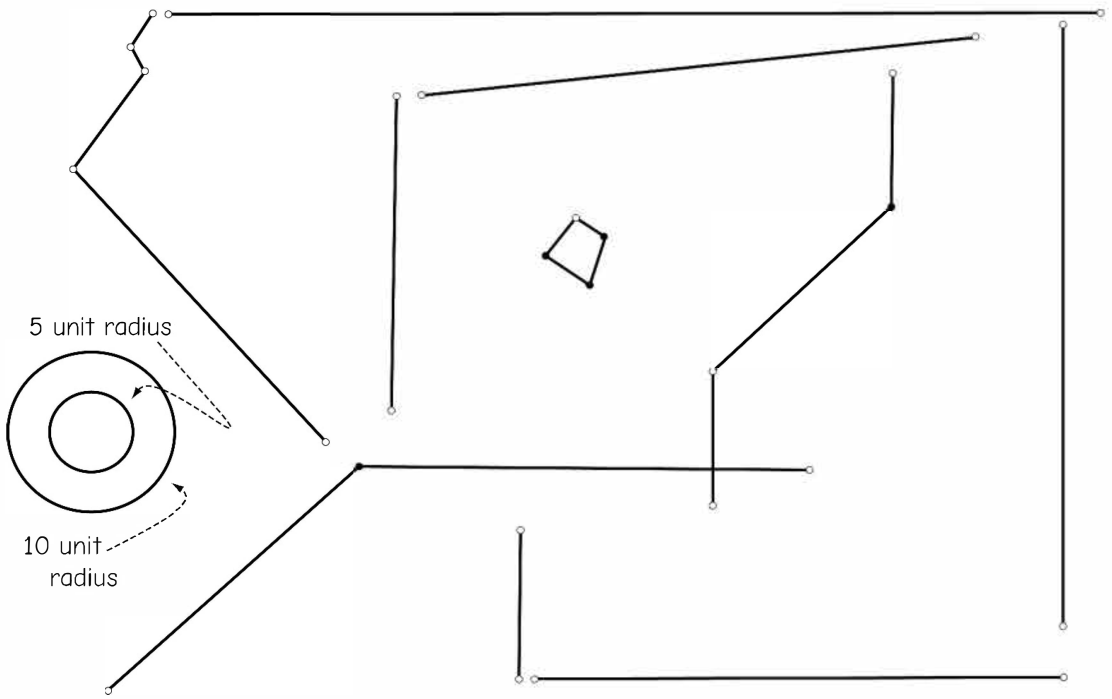
4.12 - Identify a characteristic feature or error in digitizing at each of the labeled letter locations in the drawing below; for example, node, overshoot, missing label, etc.:



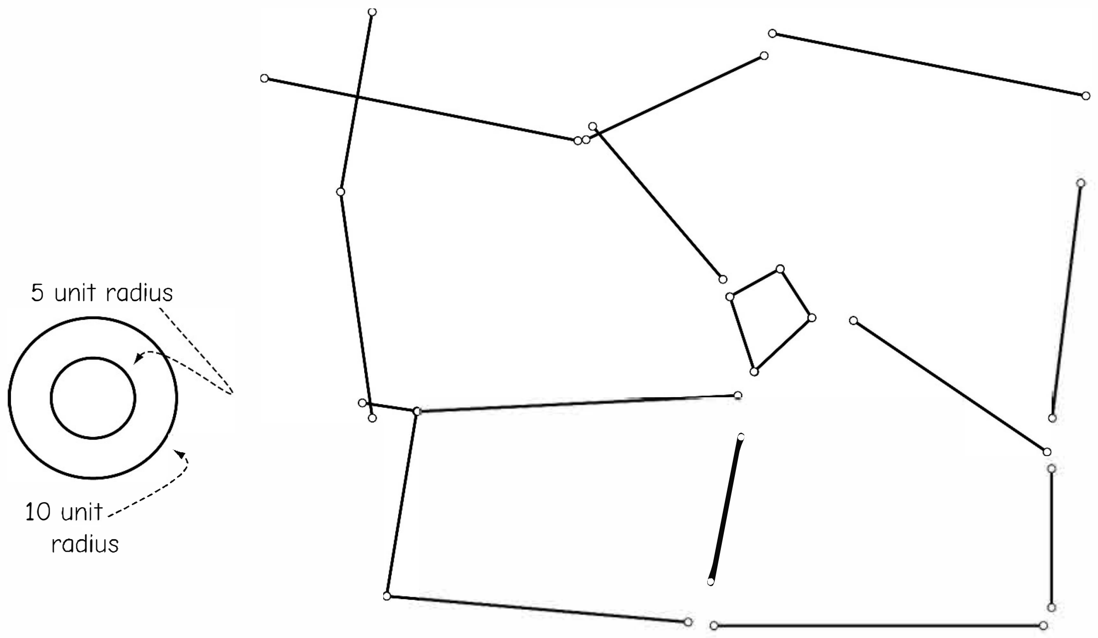
4.13 - Identify a characteristic feature or error in digitizing at each of the labeled letter locations in the drawing below; for example, node, overshoot, missing label, etc.:



4.14 - Sketch the results of combined node (open circle), vertex (closed circle), and edge (lines) snapping with a snap tolerance of a) a distance of 5 units, and b) a distance of 10 units, as shown by the snap circles. Note the radius and not the diameter of these circles defines the snapping distance.



4.15 - Sketch the results of combined node (open circle), vertex (closed circle), and edge (lines) snapping with a snap tolerance of a) a distance of 5 units, and b) a distance of 10 units, as shown by the snap circles. Note the radius and not the diameter of these circles defines the snapping distance.



4.16 - What are splines, and how are they used during digitizing?

4.17 - a) Why is line thinning sometimes necessary?

b) Does increasing the width of the line thinning band tend to increase, decrease, or not affect the number of points removed?

c) Does increasing the number of points initially spanned tend to increase, decrease, or not affect the number of points removed?

4.18 - Contrast manual digitizing to the various forms of scan digitizing. What are the advantages and disadvantages of each?

4.19 - What is the "common feature problem" when digitizing, and how might it be overcome?

4.20 - Describe the general goal and process of map registration.

4.21 - What are control points, and where do they come from?

4.22 - Define an affine transformation, including the form of the equation. Why is it called a linear transformation?

4.23 - What is the root mean square error (RMSE), and how does it relate to a coordinate transformation?

4.24 - Is the average positional error likely to be larger, smaller, or about equal to the RMSE? Why?

4.25 - Why are higher-order (polynomial) projections to be avoided under most circumstances?

4.26 - Which of the following transformations will likely have the smallest average error at a set of independent test points?

a) affine, RMSE = 14.23
b) affine, RMSE = 9.8
c) 2nd-order polynomial, RMSE = 9.7
d) 3rd-order polynomial, RMSE = 6.45

4.27 - Which of the following transformations will likely have the smallest average error at a set of independent test points?

a) 1st-order polynomial, RMSE = 5.3
b) affine, RMSE = 9.8
c) 2nd-order polynomial, RMSE = 4.9
d) 1st-order polynomial, RMSE = 9.9

4.28 - Define and describe metadata. Why are metadata important?

5 Global Navigation Satellite Systems and Coordinate Surveying

Introduction

Broadly defined, there are two general ways we measure the locations of geographic features. The first uses field measurements, and is described in this chapter. We travel to a feature and physically occupy a location to measure unknown X, Y, and often Z coordinates. Measurement systems have become quite sophisticated, incorporating satellite and laser technologies, primarily Global Navigation Satellite Systems (GNSS), as well as traditional ground surveying methods. Field measurements may be accurate to within millimeters (tenths of inches).

The second set of location measurement techniques uses remote data collection, primarily from aerial and satellite images. Coordinate positions may be obtained to within a few centimeters (inches) from properly collected, carefully processed images. Image systems are described in Chapter 6.

GNSS are satellite-based technologies that give precise positional information, day or night, in most weather and terrain conditions (Figure 5-1). GNSS technologies may help navigate and track moving objects large enough to carry a receiver. Receivers shrink in size, weight, and power requirements each year.

GPS, for Global Positioning System, is sometimes used synonymously for GNSS. GPS more specifically refers to the U.S. developed satellite navigation system, the first developed and deployed globally.

Figure 5-1: An artist's rendering of a Galileo satellite, part of a planned 30-satellite constellation at the heart of a European Union led satellite navigation system (courtesy Lockheed Martin).

Coordinate surveying is often used to complement GNSS measurements. Coordinate surveying encompasses optical and electronic angle and distance measurements, some already described in Chapter 3. Because both GNSS and coordinate surveying measurements are important, they will be covered in this chapter, first by describing GNSS and coordinate surveying tools and methods, and then discussing common applications.

GNSS Basics

Because they are inexpensive, accurate, and easy to use, GNSS have had a pervasive impact in the geographic information sciences. GNSS have become the most common method for field data collection in GIS.

As of 2019 there are three functioning GNSS systems, a fourth very near completion, and two regional systems under development. The U.S. NAVSTAR Global Positioning System (GPS) was the first deployed and is the most widely used system. There is an operational Russian system named GLONASS with 24-hour coverage worldwide. A third system includes a constellation of 30 positioning satellites with global coverage, the Chinese Compass, or BeiDou, Satellite Navigation System. A fourth system, Galileo, is being developed by a consortium of European governments and industries, with a planned total of 30 satellites in the constellation, scheduled for completion in 2020. There is a regional system by the Indian government (IRNSS), with seven satellites giving coverage to south-central Asia, operational in 2016, but semi-functional because of equipment failures. A regional system, QZSS, is under development for Japan, east Asia, and the western Pacific Ocean. In the following discussion we use GNSS as a generic term for all four global systems, and use GPS to refer specifically to the U.S. NAVSTAR system.

There are three main components, or segments, of any GNSS (Figure 5-2). The first is the *satellite segment.* This is a constellation of satellites orbiting the Earth and transmitting positioning signals. The second component of any GNSS is a *control segment*. This includes tracking, communications, data gathering, integration, analysis, and control facilities. The third part of GNSS is the *user segment*, the GNSS receivers.

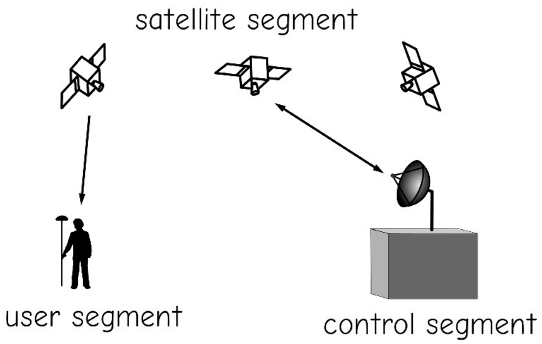


Figure 5-2: The three segments that comprise a GNSS.

A *GNSS receiver* is a device that records data transmitted by each satellite, and then processes these data to obtain three-dimensional coordinates (Figure 5-3). There is a wide array of receivers and methods for determining position. Receivers are often handheld devices with screens and keyboards, or electronic components mounted on cars and trucks, planes, or other objects.

The satellite and control segments differ for each GNSS. The NAVSTAR GPS includes a constellation of satellites orbiting the Earth at an altitude of approximately 20,000 km. Initial system design included 21 active GPS satellites and three spares, distributed among six offset orbital planes. Every satellite orbits the Earth twice daily, and each satellite is usually above the flat horizon for eight or more hours each day. Experimental and successive operational satellites have outlasted their design life, so there have typically been more than 24 satellites in orbit simultaneously. Between four to eight active satellites are typically visible from any unobstructed viewing location on Earth.

GPS is controlled by a set of ground stations. These are used to observe, maintain, and manage satellites, communications, and related systems. There are five tracking stations in the GPS system, spread across the planet. Data are gathered from a number of sources by the stations, including satellite health and status from each GPS satellite, tracking information from each tracking station, timing data from the U.S. Naval Observatory, and surface data from the U.S. Defense Mapping Agency. A Master Control Station synthesizes information and broadcasts navigation, timing, and other data to each satellite. The Master Control Station also signals each satellite as appropriate for

course corrections, changes in operation, or other maintenance.

The GLONASS system is another currently operating GNSS. GLONASS was initiated by the former Soviet Union in the early 1970s. Satellites were first launched in the early 1980s, and the system became functional in the mid-1990s. The GLONASS system was designed for military navigation, targeting, and tracking, and is operated by the Russian Ministry of Defense, with control and tracking stations similar to those for the NAVSTAR GPS system.

GLONASS was designed to include 21 active satellites and three spares. New designs have been phased in as older satellites have expired, and system managers have focused on maximizing coverage over Russia. The GLONASS system is established, with a published renovation and maintenance plan, such that commercial manufacturers have developed dual GPS/GLONASS capable receivers.

The Chinese Compass (BeiDou) system consists of 30 satellites with attendant ground station infrastructure. Eighteen satellites were launched in 2018, and 10 upgraded satellites are scheduled for launch in 2019. The constellation includes both geostationary and inclined orbiting satellites. The system is designed for both civilian and military use, with substantially augmented accuracies, up to a few centimeters within China, due to a network of local receivers.

The European Galileo system also implements satellite, control, and user segments. There are 30 satellites planned for the complete Galileo constellation. Satellites are arranged in three orbital paths at a 54° orbital inclination, with a satellite altitude near 23,600 km above the Earth. This satellite constellation will provide better coverage of high northern latitudes than the U.S. NAVSTAR GPS system, to better serve northern Europe. Galileo will be managed through two control centers in Europe and 20 Galileo Sensor Stations spread throughout the world to monitor, communicate with, and relay information among satellites and the control centers.

Figure 5-3: A hand-held GNSS receiver (left) and a GNSS receiver in use (right, courtesy Juniper Systems).

GNSS Broadcast Signals

GNSS positioning is based on radio signals broadcast by each satellite. Systems vary, but all systems transmit on multiple frequencies, and also send additional data needed to calculate positions from the main signals. As an example, the NAVSTAR GPS satellites broadcast positioning signals on three base frequencies (Table 5-1), the L1, L2, and L5. These *carrier signals* are modulated to produce *coded signals*, e.g., the C/A code at 1.023 MHz and the P and M code at 10.23 MHz. The L1 signal carries both the C/A and P codes, while the L2 carries the P and M codes (Figure 5-4) Additional signals provide information, and coded signals are sent containing satellite navigation and other information. These other signals have been added to improve function, for example, the CNAV for tracking, a forward error correction code, and and MNAV for enhanced military applications. The coded signals (C/A, P, and M) are sometimes referred to as the *pseudorandom code*, because they appear quite similar to random noise. However, short segments of the code are unique for each satellite and time. A receiver decodes each signal to identify the satellite, transmission time, and satellite position at the time the signal was sent. The receiver combines this information from multiple satellites for positioning. The coded signal does repeat, but the repeat interval is long enough to not cause problems in positioning.

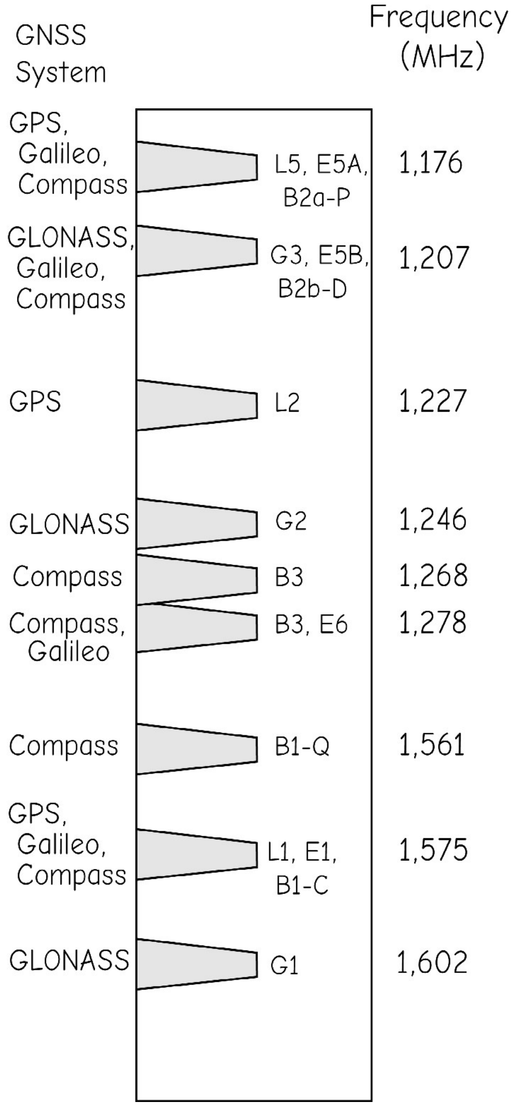


Figure 5-4: Existing and proposed GNSS broadcast signals, frequencies, and positioning services. Signals are spaced to avoid interference, or coded where they overlap. Frequencies are not spaced to scale (courtesy ESA).

Table 5-1 : GPS Signals

Name	Frequency (MHz)
L1, L1C	1,575.42
L2, L2CM, L2CL	1,227.6
L5	1,176.45
P, M	10.23
C/A	1.023

.Positions based on carrier signal measurements (L1, L2, and L5 frequencies for the NAVSTAR GPS), and positions based on multiple frequencies are inherently more accurate than those based on the code signal or single frequency measurements. The mathematics and physics of carrier measurement are better suited for making positional measurements. Comparing two frequencies may improve error removal, primarily by removing something called ionospheric errors, described later.

Improved accuracy in GNSS positioning usually incurs an added cost, in more expensive equipment or in time spent collecting and processing data. Carrier measurements require more sophisticated and expensive receivers and must record signals for longer periods of time than code receivers. If the signal is blocked by a building, mountain, or other object, the signal may be lost momentarily and carrier phase measurements begun anew. This substantially reduces the efficiency of carrier phase data collection, although these constraints have decreased with modern receivers. Newer systems are often capable of tracking multiple GNSS constellations with hundreds of channels, reducing loss of lock.

GNSS satellite also broadcasts data on satellite status and location. Data parts go by various names, but using the GPS conventions, the information includes an *almanac*, data used to determine the satellite status, and *ephemeris data* of satellite tracks. These ephemerides allow a GNSS receiver to accurately calculate the position of the broadcasting satellite and the expected positions of other satellites. Satellite health, clock corrections, and other data are also transmitted.

The various GNSS systems span a similar range of frequencies, and are organized so that there is little interference between any two signals, even when they share the same fundamental frequency (Figure 5-4). GLONASS broadcasts G1 and G2 carriers similar to GPS, and an additional G1/G10 and experimental G3 signal at higher and lower frequencies. BeiDou/Compass and Galileo include the broadcasts of a range of signals on several fundamental carriers, including overlaps with the GPS, BeiDou/Compass, and GLONASS signals at various frequencies.

Substantial effort has been directed at ensuring the NAVSTAR, GLONASS, Compass/BeiDou, and Galileo systems do not interfere with each other, and are as compatible as practical. This simplifies the production of dual-system receivers that are capable of using multiple GNSS signals, improving coverage, accuracy, and efficiency. When all GNSS constellations are fully operational, there may be as many as 70 satellites broadcasting in this range, greatly improving overall utility.

Range Distances

GNSS positioning is based primarily on *range distances*. A range distance, or *range*, is a distance between two objects. For GNSS, the range is the distance between a satellite in space and a receiver (Figure 5-5). GNSS signals travel approximately at the speed of light. The range distance from the receiver to each satellite is calculated based on signal travel time from the satellite to the receiver:

$$\text{Range} = \text{speed of light} * \text{travel time} \quad (5.1)$$

Coded signals are used to calculate signal travel time by matching sections of the code. Timing information is sent with the coded signal, allowing the GNSS receiver to calculate the precise transmission time for each code fragment. The GNSS receiver also observes the reception time for each code fragment. The differences between transmission and reception times are used to calculate range distances, often at rates as high as

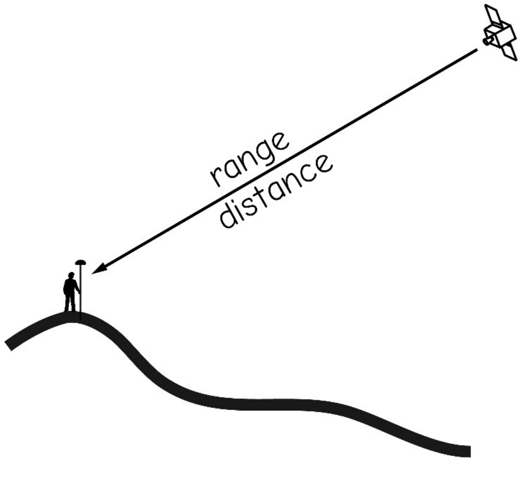


Figure 5-5: A single satellite range measurement.

one new range calculation a second (Figure 5-6).

Carrier phase GNSS is also based on a set of range measurements. In contrast to coded signals, the phase of the satellite signal is measured. Each individual wave transmitted at a given frequency is identical, and at any given point in time there is some unknown integer number of waves plus a partial wave that fit in the distance between the satellite and the receiver. Carrier signal observations over time intervals allow the calculation of wavelength number over the measurement interval, and then the calculation of very precise satellite ranges.

Simultaneous range measurements from multiple satellites are used to estimate a receiver's location. A range distance and a satellite position define a sphere, so measurement of a range distance to one satellite places the receivers somewhere on that sphere (Figure 5-7a). Range measurements from two satellites place the receiver on a circle formed by the intersection of two spheres (Figure 5-7b). Range measurements from three satellites define three spheres that intersect at two points (Figure 5-7c). A sequence of range measurements through time from three satellites will reveal that one of the points remains nearly stationary, while the other point moves rapidly through space. The point moves because the size and relative geometry of the spheres change through time as the satellites change positions. Simultaneous measurements from four or more satellites (Figure 5-7d) are usually required to reduce receiver clock errors and to allow instantaneous position measurement with a moving receiver. Data collected from more than four satellites usually improves the accuracy of position measurements.

Positional Uncertainty

Errors in range measurements and uncertainties in satellite location introduce errors into GNSS-determined positions (Figure 5-8). Range errors vary substantially even if range measurements are taken just a few seconds apart. Errors in the ephemeris data lead to erroneous estimates of the satellite position, causing location error. Clock, atmospheric, and ionospheric uncertainties add error to range measurements, resulting in a band of range uncertainty around the GNSS receiver position.

Several methods help us improve our measurement accuracy. Point averaging is perhaps the simplest and most common. We may collect many position fixes on a stationary receiver. Most receivers may estimate a new position, or fix, every second. Averaging also yields a cluster of individual fixes distributed about the mean location. The dispersion of the cluster may help quantify positional accuracy.

Multiple position fixes are not possible when collecting data while moving, for example, when determining the location of an airborne plane. Also, averaging does not

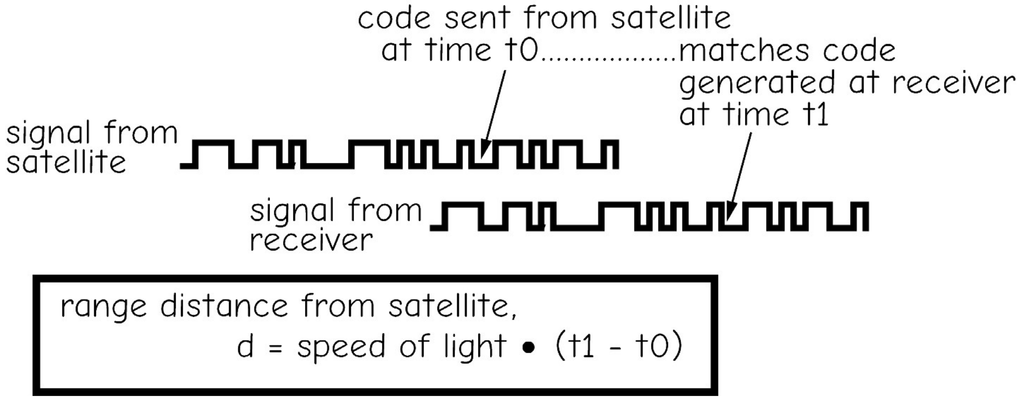


Figure 5-6: A decoded C/A satellite signal provides a range measurement.

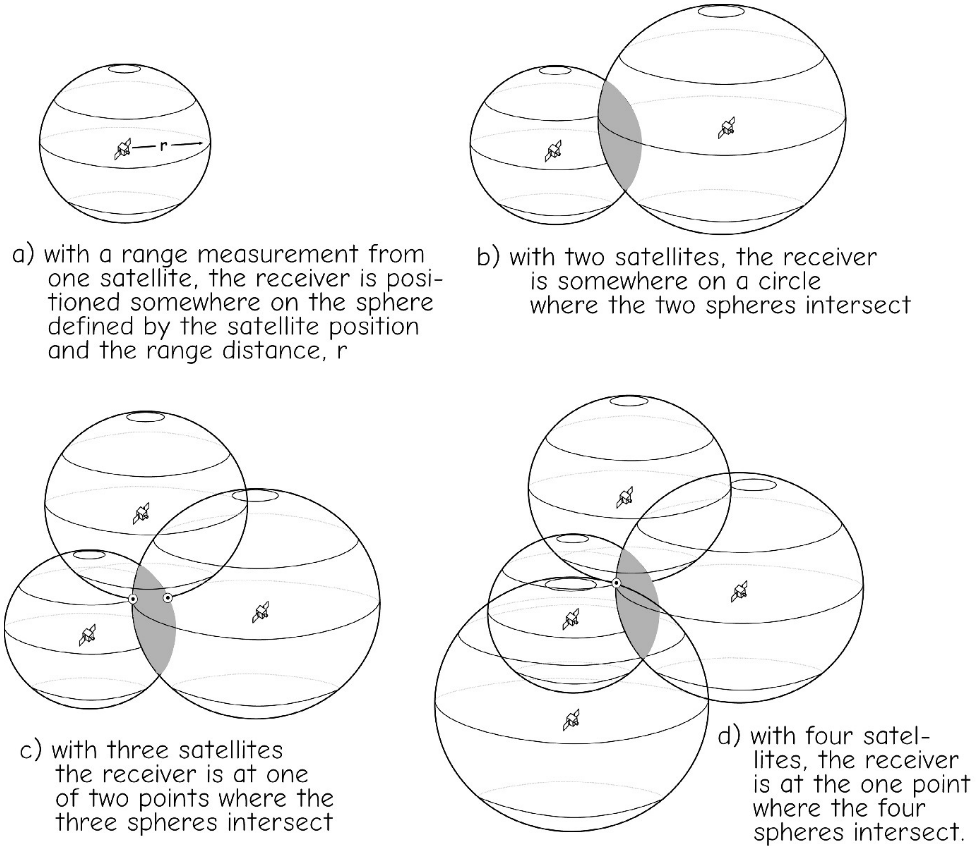


Figure 5-7: Range measurements from multiple GNSS satellites. Range measurements are combined to narrow down the position of a GNSS receiver. Range measurements from more than four satellites may be used to improve the accuracy of a measured position (adapted from Hurn, 1989).

remove any bias in the calculated position. Alternative methods for reducing positional error rely on reducing the several sources of range errors.

Sources of Range Error

Ionospheric and *atmospheric delays* are common sources of GNSS range error. Range calculations depend on the speed of light. While this speed is constant when light is passing through a uniform electromagnetic field and in a vacuum, these conditions don't hold for GNSS signals. The Earth is surrounded by a varying density of charged particles in the ionosphere, formed by incoming solar radiation, which strips electrons from elements in the upper atmosphere. Changes in the charged particle density may affect GNSS transmissions.

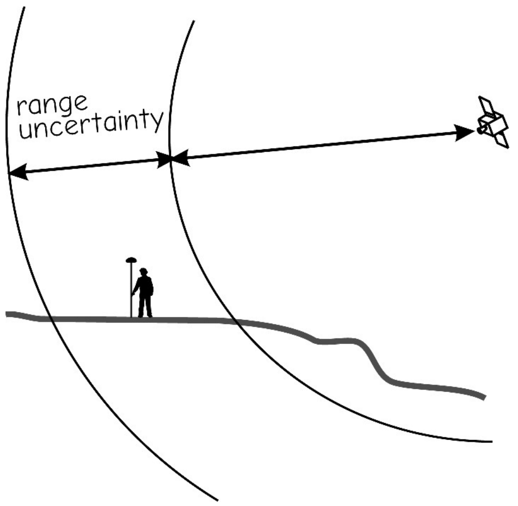


Figure 5-8: Uncertainty in range measurements leads to positional errors in GNSS measurements.

Atmospheric density is significantly different from that of a vacuum. Density variation is due largely to changes in temperature, atmospheric pressure, and water vapor. Range errors occur because the GNSS signal velocity changes as it passes through the ionosphere and atmosphere; some systems allow satellite screening based on horizon angle, to reduce atmospheric path effects on accuracy (Figure 5-9).

While we may attempt to reduce ionospheric errors by adjusting for changes in the speed of light, this is rarely possible. The electromagnetic field varies both in time and space, and there is no practical way to measure this variation reliably for rapid correction. Analytical correction is typically reserved for very precise surveys.

Errors can be reduced by receiver design, because ionospheric effects depend on frequency. *Dual-frequency* receivers collect information on multiple GNSS signals simultaneously, and use sophisticated models to remove most of the ionospheric errors. Dual-frequency receivers are dropping in cost rapidly, and may soon be widely affordable. However, there is no analytical method to remove atmospheric range errors. These are best removed by differential correction, described in the next section.

Range errors come from system operation and delays. Satellite tracking is imperfect, and timing and other signals are slowed during transmission through the entire system. Atomic clocks on the satellite may be in error, although these are typically small. Many of these errors may be partially removed in rigorous analytical post-processing, rarely applied due to the complexity of the calculations and needed additional data. Differential correction, described later, also removes much of the systemic range error.

Receivers also introduce errors into GNSS positions. Receiver clocks may contain biases. Signals may reflect off of objects prior to reaching the antenna. These reflected, *multipath* signals have a longer, erroneous range than direct GNSS signals. Multipath signals often have lower power, and so may be screened by setting a threshold signal-to-noise ratio. Signals with high noise relative to the mean signal strength are ignored. Multipath signals may also be screened by properly designed antennas. Multipath signals are most commonly a problem in urban settings that have an abundance of strong corner reflectors, such as the sides of buildings and streets. This is often the largest source of range error, particularly when collecting data without a specialized multi-path rejection antenna.

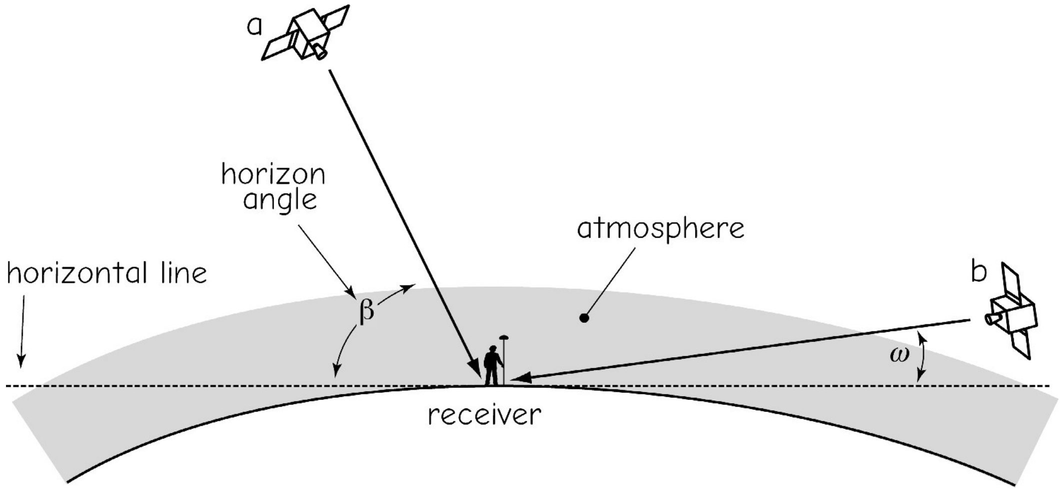


Figure 5-9: GNSS receivers often discard signals from satellites near the horizon. As this image shows, signals from satellites high above the horizon (a) with high horizon angles (β) have shorter path lengths in the atmosphere than low-angle satellites (b, with angle ω). Atmospheric delays and hence range errors are larger for satellites with low horizon angles. Typically, a “mask” is set at approximately 15 degrees above the horizon, and satellites are ignored if they are below this limit.

Satellite Geometry and Dilution of Precision

The geometry of the GNSS satellite constellation is another factor that affects positional error. Range errors create an area of uncertainty perpendicular to the transmission direction of the GNSS signal. These areas of uncertainty may be visualized as a set of nested spheres, with the true position somewhere within the volume defined by the intersection of these spheres (Figure 5-10). These areas of uncertainty intersect, and the smaller the intersection area, the more accurate the position fixes are likely to be. Signals from widely spaced satellites are complementary because they result in a smaller area of uncertainty. Signals from satellites in close proximity overlap over broad areas, resulting in large areas of positional uncertainty. Widespread satellite constellations provide more accurate GNSS position measurements.

Satellite geometry is summarized in a number called the *Dilution of Precision*, or DOP (Figure 5-11). There are various kinds of DOPs, including the Horizontal (HDOP), Vertical (VDOP), and Positional (PDOP) Dilution of Precision. The PDOP is most used and is the ratio of the volume of a tetrahedron created by the four most widespread, observed satellites to the volume defined by the ideal tetrahedron. This ideal tetrahedron is formed by one satellite overhead and three satellites spaced at 120-degree intervals around the horizon. This constellation is assigned a PDOP of one, and closer groupings of satellites have higher PDOPs. Lower PDOPs are better. Most GNSS receivers review the almanac transmitted by the GNSS satellites and attempt to obtain measurements that include the constellation with the lowest PDOP. If this best constellation is not available, for example, some satellites are not visible, successively poorer constellations are tested until the best available constellation is found. The receivers typically provide a measurement of PDOP while data are collected, and a maximum PDOP threshold may be specified, above which data are not collected.

Range errors and DOPs combine to affect GNSS position accuracies. There are many sources of range error, and these combine to form an overall range uncertainty for the measurement from each visible GNSS

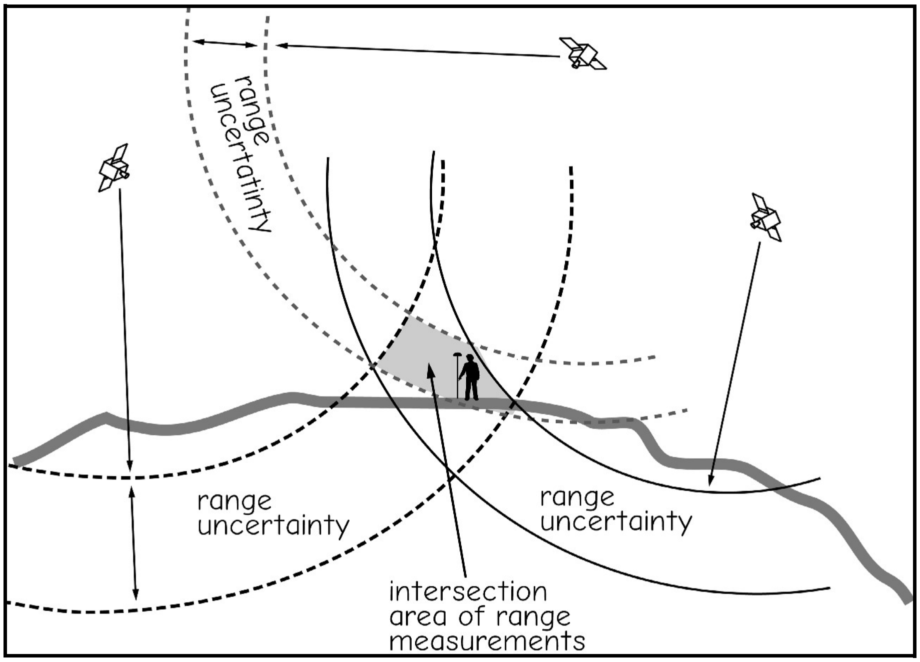


Figure 5-10: Relative GPS satellite position affects positional accuracy. Range uncertainties are associated with each range measurement. These combine to form an area of uncertainty at the intersection of the range measurements.

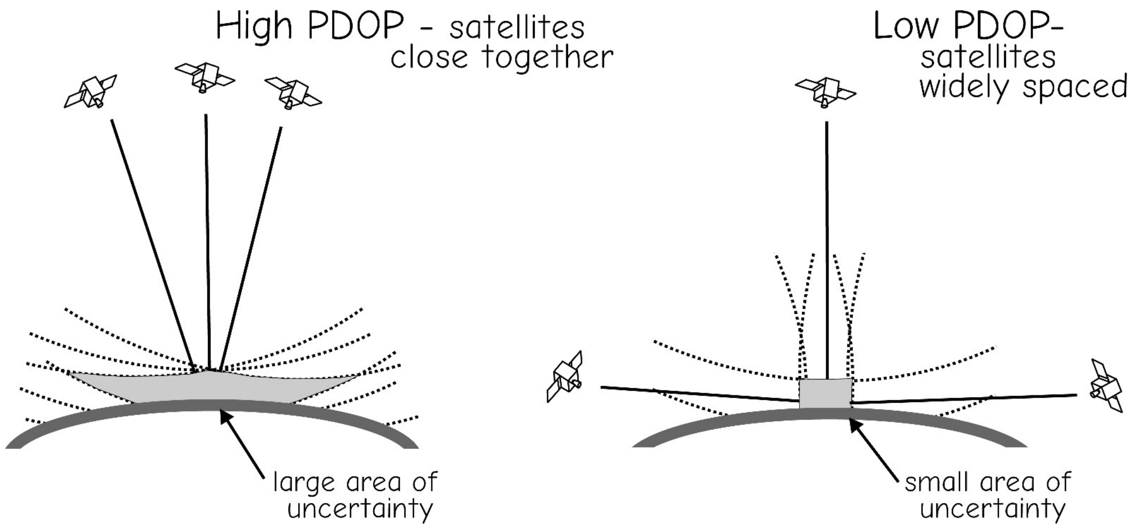


Figure 5-11: GPS satellite distribution affects positional accuracy. Closely spaced satellites result in larger positional errors than widely spaced satellites. Satellite geometry is summarized by PDOP, with lower PDOPs indicating better satellite geometries.

satellite. If more precise coordinate locations are required, then the choices are to use equipment that makes more precise range measurements, and/or to collect data when DOPs are low.

GNSS accuracies depend on the type of receiver, atmospheric and ionospheric conditions, the number of range measurements, the satellite constellation, and the algorithms used for position determination (Figure 5-12). Current C/A code receivers typically provide accuracies between 3 and 30 m for a single fix. Errors larger than 100 m for a single fix occur occasionally. Accuracies may be improved substantially, to between 2 and 15 m, when multiple fixes are averaged. The longer the data collection time, the greater the accuracy. Improvements come largely from reducing the impact of rarer, large errors, but average accuracies are rarely below 1 m when using a single C/A code receiver.

Accuracies when using carrier phase, dual frequency, or similar receivers are much higher, on the order of a few centimeters. Ionospheric effects vary by wavelength. Effects can be compared in dual frequency receivers, and errors partially removed. These accuracies come at the cost of longer data collection times, although clever analysis and system design have lowered this to a few to tens of minutes, rather than hours as in the recent past. Fast, inexpensive, dual-frequency chipsets are becoming available that may provide 20 cm (eight inch), real-time accuracy. Another technique, differential correction, is the most reliable means of obtaining 30 cm (one foot) accuracy. This technique is described in the following section.

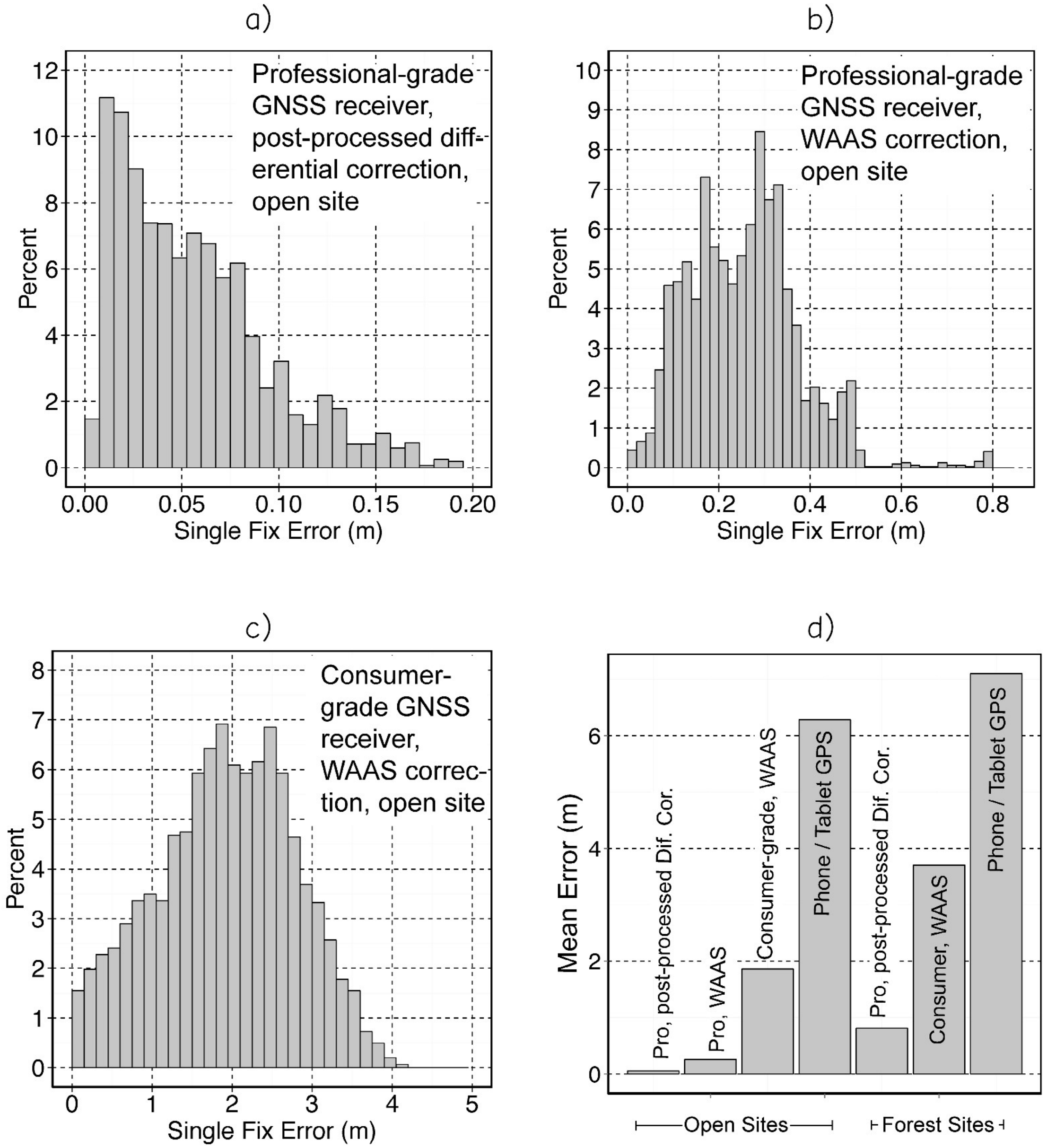


Figure 5-12. Observed GPS error distributions for various receivers under open sky conditions (a through c) and mean error under open sky and dense deciduous forest canopy (d). Results show highest accuracies for a professional-grade GNSS receiver (panel a, TRIMBLE 6H, post-processed carrier phase differential correction, < 20 km to a base station), a professional-grade GNSS receiver with WAAS real-time correction (b), and an inexpensive consumer-grade GNSS (c, Garmin etrex). Mean errors for both open sky and under forest canopy are shown in panel d for these three receivers/configurations, and for an iPhone 5 and a similar tablet GPS. Centimeter-level accuracies are available with the best equipment under optimal conditions, but less expensive receivers and obstructed skies reduce accuracies. In some instances, the individual fix error is important, for example, when digitizing lines or polygon boundaries, while average errors may be more important when collecting point features, allowing multiple fixes. Note that errors decrease as technology improves, but the higher-priced receivers may not be more expensive over the life of a project if lower accuracies require significant manual editing for GNSS collected data (courtesy Andy Jenks).

Differential Correction

The previous sections have focused on GNSS position measurements collected with a single receiver. This operating mode is known as autonomous GNSS positioning. An alternative method, known as *differential positioning*, employs two or more receivers. Differential positioning measurements are used primarily to remove most of the range errors and thus greatly improve the accuracy of GNSS positions (Figure 5-12). However, differential positioning is not always employed, because single receiver positioning is accurate enough for some applications, and differential positioning requires more time and/or greater expense.

Differential GNSS positioning entails establishing at least one independent *base station* receiver at a known coordinate location (Figure 5-13). The true coordinate location of the base station is typically determined using high-accuracy surveying methods, for example, repeated astronomical observations, highest-accuracy GNSS, or precise ground surveys, as described in Chapter 3.

We use the base station to estimate range measurement errors for each position fix. Remember that GNSS is based on a set of range measurements, and these range measurements contain errors. Some of these errors are due to uncertainty in the measured travel times from the satellite to the receiver. These combined travel time errors, also known as timing errors, are often among the largest sources of positional uncertainty.

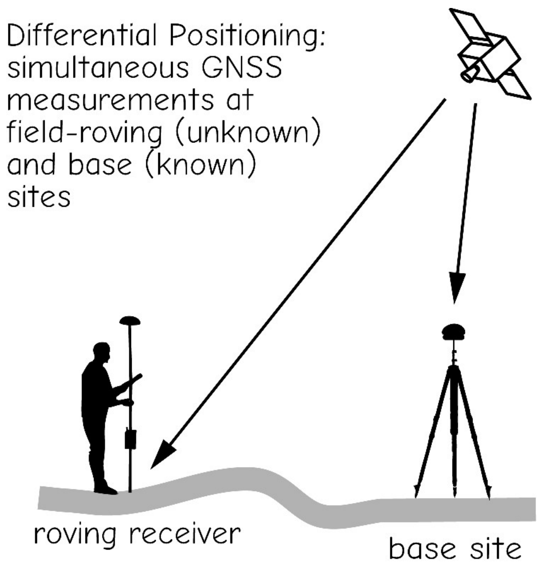


Figure 5-13: Differential GNSS positioning.

In differential correction, we use the known base station position to estimate the timing errors and hence range errors. Each GNSS satellite broadcasts its position along with the ranging signal. The "true" distance from a given satellite to the base station can be calculated because the base station and satellite locations are known. However, note the qualifying quote marks around the true distance. We cannot exactly define where the satellite is, and the base station coordinates have some (usually small) level of uncertainty associated with them. However, if we are very careful about surveying the location of our base station, then the errors in the base-to-satellite measurement are almost always smaller than the range errors contained in our uncorrected timing measurement.

The difference between the true distance and GNSS-measured distance is used to estimate the timing error for a given satellite at any given time. The timing errors change each second, so they should be measured frequently.

Timing corrections may be applied to the range measurements collected by a roving receiver (Figure 5-14). These roving receivers are used to measure GNSS positions at field locations with unknown coordinates. The timing error, and hence range error, for each satellite observed at a field location is assumed to be the same as the range error observed simultaneously at the base station. We adjust the timing of each satellite measurement made by the rover, then calculate the rover's position in the field. This adjustment usually reduces each range error and substantially improves each

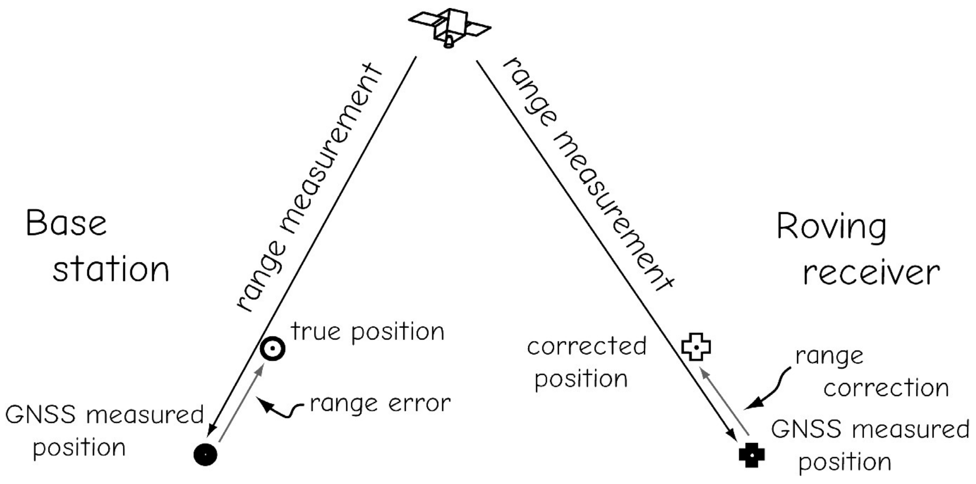


Figure 5-14: Differential correction is based on measuring GNSS timing and range errors at a base station, and applying these errors as corrections to simultaneously measured rover positions.

position fix taken with the roving field receivers.

The timing errors change across the surface of the Earth, and this places a restriction on the use of differential GNSS correction. Our roving receivers must be "near" our base station for differential correction to work. A substantial portion of the range error is due to atmospheric and ionospheric interference with the GNSS signal. Fortunately, these conditions often vary slowly with distance through the atmosphere, so interaction in one location is likely to be similar to interaction, and thus error, in a nearby location. Therefore, as long as the rover is close to the base station, within a few tens to hundreds of kilometers, we may expect differential correction to improve our position measurements.

Differential correction requires the base station and roving receivers collect data from a similar set of satellites. We cannot fix a timing error we do not measure. Any four satellites providing acceptable PDOPs will suffice, although more are better.

Successful differential correction also requires near simultaneity in the base and rover measurements. Errors change rapidly through time. If the base and rover measurements are collected more than a few tens of seconds apart, they do not correspond to the same set of errors, and thus the difference at the base station cannot be used to correct the rover data. Many systems allow data collection to be synchronized to a standard timing signal, thereby ensuring a good match when the error vectors are applied to correct the roving receiver GNSS data.

Base station data and roving receiver data must be combined for differential correction. A base station correction may be calculated for each fix and then applied to the roving receiver data. Data are often stored, downloaded from both receivers, and combined on a computer. Software provided by most GNSS system vendors is then used to compute and apply the differential corrections to the position fixes. This is known as post-processed differential correction, as corrections are applied after, or post, data collection (Figure 5-15, top).

Post-processed differential positioning is appropriate for many projects. Road locations may be digitized with a GNSS receiver mounted to the top of a vehicle. The vehicle is driven over the roads to be digitized, the rover data differential corrected, and then exported as a data layer suitable for GIS.

Post-processed differential positioning has one serious limitation. Because precise positions are not known when the rover is in the field, post-processing technologies are useless for precise navigation. A surveyor recovering buried or hidden property corners

often needs to navigate to within a few tens of centimeters of a position while in the field, so that monuments, stakes, or other markers may be recovered. When using post-processed differential GNSS, the field receiver is operating as an autonomous positioning device, and accuracies of a few meters to tens of meters are expected. This is not acceptable for many navigation purposes because too much time will be spent searching for the final location.

Real-time Differential Positioning

An alternative GNSS correction method, known as *real-time differential correction*, may be appropriate when precise navigation is required. real-time differential correction requires some extra equipment and there is some cost in slightly lower accuracy when compared to post-processed differential GNSS. However, the accuracy of real-time differential correction is substantially better than autonomous GNSS, and accurate locations are determined while still in the field.

Real-time differential GNSS positioning requires a communications link between base stations and the roving receiver (Figure 5-15, bottom). Typically, the base station is connected to a radio transmitter and an antenna. FM radio links are often used due to their longer range and good transmission through vegetation and other obstacles, or cell phone networks are also often used. The base station collects a GNSS signal and calculates range distances. The error is calculated for each range distance. The magnitude and direction of each error is passed to the radio transmitter, along with information on the timing and satellite constellation used.

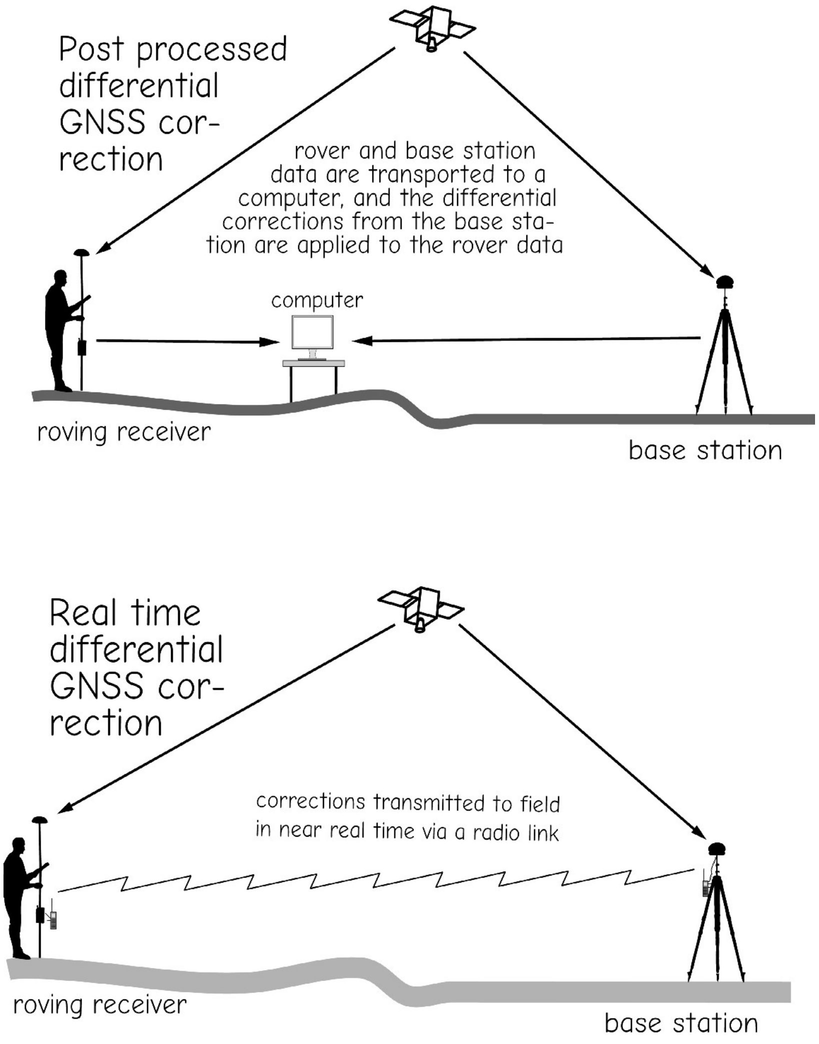


Figure 5-15: Post-processed and real-time differential GNSS correction.

This continuous stream of corrections is broadcast via the base station radio and antenna.

Roving GNSS receivers are outfitted with a radio, cell phone, or other communication system, and any receiver within the broadcast range of the base station may receive the correction signal. The roving receiver is also recording GNSS data and calculating position fixes. Each position fix by the roving receiver is matched to the corresponding correction from the base station. The appropriate correction is then applied to each fix and accurate field locations are computed in real time.

Real-time differential correction requires a broadcasting base station; however, every user is not required to establish a base station and complete communication system. For example, the U.S. Coast Guard has established a set of GPS *radio beacons* in North America that broadcast a standardized correction signal (Figure 5-16). A compatible GPS receiver near these beacons can use the signal for differential correction. These GPS *beacon receivers* typically have an additional antenna and electronics for processing the beacon signal. Beacons were originally placed to aid ship navigation in coastal and major inland waters, so beacons are concentrated near coastal and major inland waters. This system has become the Maritime Differential GPS system and is part of a National Differential GPS system (NDGPS) under development with collaboration of Federal Departments of Transportation, Homeland Security, and others. NDGPS will support navigation and positioning in areas distant from the Coast Guard network. Many GPS manufacturers sell beacon receiver packages that support real-time correction using the beacon signal.

WAAS, Augmentation, and Satellite-based Corrections

There are alternatives to ground-based differential correction for improving the accuracy of GNSS observations. One alternative, known as the Wide Area Augmentation System (WAAS), is administered by the U.S. Federal Aviation Administration to provide accurate, dependable aircraft navigation.

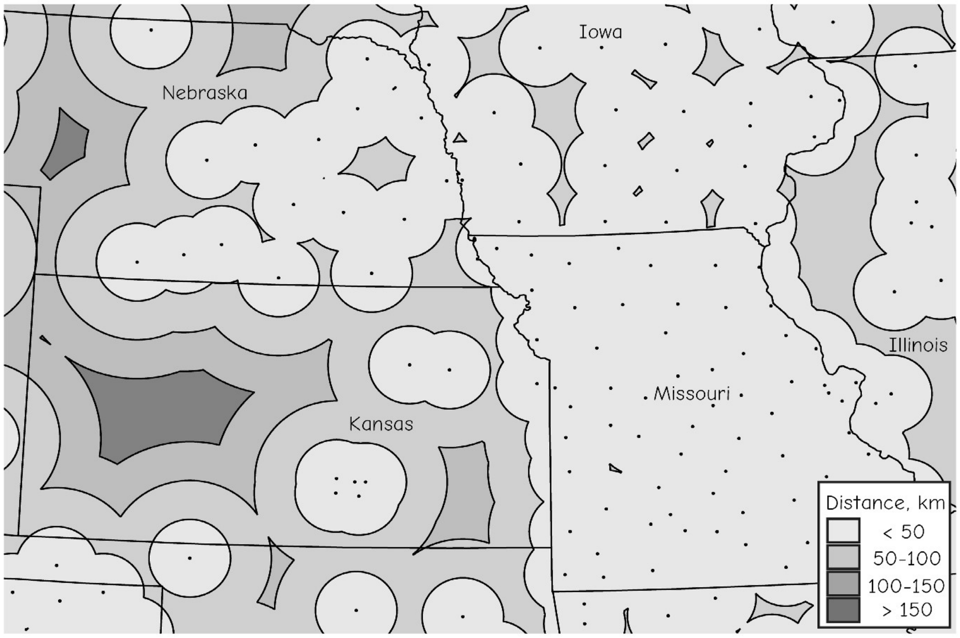


Figure 5-16: The location of radio beacon stations (dots) in the central U.S., in April, 2015. Distances from the nearest beacon are shown in various shades of gray.

WAAS uses a network of ground reference stations spread across North America to correct GPS signals. A generalized correction for each station is broadcast from geostationary satellites, and applied in real time for improved accuracy in roving receivers. Tests indicate individual errors are less than 7 m 95% of the time, and average errors less than 3 m, an improvement over uncorrected C/A code (Figure 5-12). WAAS is often unavailable at extreme northern latitudes, where equatorial geostationary satellites are often not visible.

Real-Time Kinematic and Virtual Reference Stations

The highest accuracy differential correction is provided with dual-frequency, carrier phase positioning, often called *real-time kinematic* (RTK) GNSS. The amount of ionospheric delay is different for different frequencies, so by comparing signals, such as the GPS carriers L1 and L2, the ionospheric delays may be estimated and removed. The CORS and WAAS differential positioning systems described so far are primarily single frequency, and so less accurate than a rigorous dual-frequency system. While single-frequency positions collected for periods of less than an hour are typically in error by tens of centimeters (a half-foot) or more, dual-frequency GNSS are often accurate to a few centimeters (an inch) or better.

RTK is such a powerful technology that many state governments are establishing a dense constellation of dual-frequency receivers in a *Virtual Reference Station* network (VRS). Stations are spaced in a network over some region such that a roving receiver is never more than an acceptable distance from a base (Figure 5-17). The systems provide dual-frequency base station data broadcast in a standard way over a given radio or cellphone signal, along with base station information. A roving dual-frequency receiver may identify the closest or best local receiver, and compare base signals to roving signals to obtain positions to within a few centimeters while in the field.

There are disadvantages to RTK GNSS. The receivers are more expensive, although prices are dropping. The roving RTK receiver must be closer to the base station for highest accuracies, typically within tens of kilometers. This requires either a denser network of base stations, or that the RTK users set up and maintain their own base station for each project. Finally, as with all carrier phase positioning, satellite signals must be continuously tracked for longer periods, although modern receivers have reduced this time to a few to tens of minutes.

Precise Point Positioning

Precise point positioning (PPP) is an alternative to differential correction. This technique uses precise satellite, clock, and orbit measurements to solve for point locations. It has the advantage of worldwide application without need of a base station, and accuracies as high as 10 cm were achieved in early incarnations of this method. Unfortunately, PPP requires com-

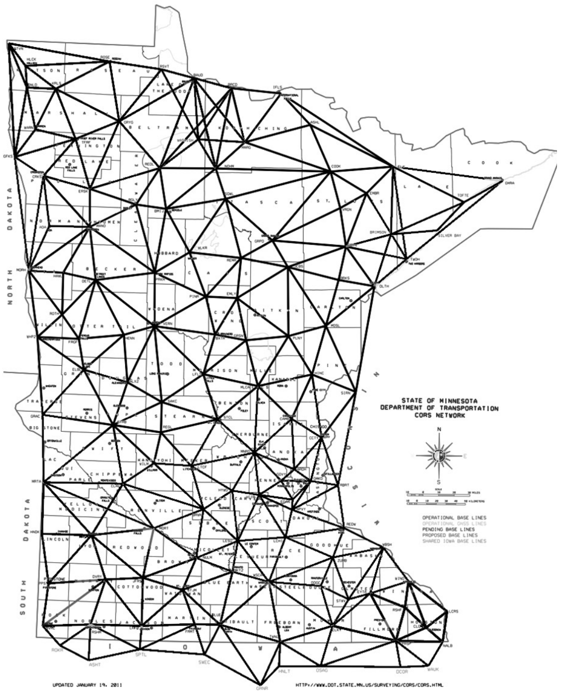


Figure 5-17: The distribution of stations in the Minnesota Department of Transportation VRS network. A station is located at each vertex in the network, ensuring close proximity for any roving receiver within the network (courtesy MNDOT).

plex calculations on long, uninterrupted observations on a satellite set for highest accuracies. Much work has been directed at integrated systems of satellite observations and communications, such that near real time precise satellite positions are calculated and used to aid in PPP-like accuracies with shorter observation times. Examples include Trimble Centerpoint-RTX, NavCom Starfire, and Veripos TERRASTAR. Typically, these systems are sold as a subscription service, in which a monthly or annual fee is paid to access the real-time satellite positioning and other data. These data may be broadcast via a cellular modem, an internet link, or a satellite radio.

A Caution on Datums

Errors may easily be injected into GNSS data due to improper datum transformations. One must be cautious in using GNSS data, either directly, or after applying a differential correction, because the datum transformation used is often not transparent, or is poorly documented. The U.S. NAVSTAR GPS system provides a good example of the confusion that may occur.

GPS satellite locations are reported in the most current WGS84 datum. Although this condition may well change with the adoption of the NATRF2022, for now the WGS84 datum is quite different from that used as the basis for GIS data in the United States. Most data are in national or local datums, for example, NAD83(2011), because there is not a dense network of points with accurate WGS84 coordinates that are readily found in the field. As noted in Chapter 3, ignoring or selecting the wrong datum transformation will introduce error into this process. GNSS vendors typically provide an option to report data in one of these commonly used coordinate systems; for example, the user may set the GNSS receiver to display UTM or State Plane coordinates, and save these to features collected in files. However, the GNSS vendors often do not clearly identify the datum transformation used. As noted in Chapter 3, early versions of the NAD83 datum that underlie the UTM system were and remain up to 2 m (6 feet) different from the WGS84 datum, so you cannot assume they are the same, as is common practice, and must carefully choose the correct datum transformation or you'll otherwise degrade your data.

Confusion may be introduced during differential correction. Here, base station coordinates define corrections relative to a point in a defined coordinate system. These coordinates may be based on a datum different from the WGS84 datum used for GPS data collection. Appropriate transformations between datums must be applied to maintain accuracy. For example, the CORS network of GPS stations is a common source of base data for differential corrections. The coordinates for these base stations were for a period typically reported in the most recent CORS realization of the NAD83 datum, but now most provide coordinates in an ITRF datum. An appropriate datum transformation must be applied when using these as a base for correction if the highest accuracies are to be maintained. As noted earlier, the differences between the later NAD83(CORS) datums and most recent realizations of WGS84 are typically less than a meter, so introduced errors may be small relative to the accuracy required for some intended analyses. However, many projects require submeter accuracy, and for some conditions the errors may be quite large, to tens or hundreds of meters, depending on the datums and projections involved. These errors may be avoided at little cost with the application of appropriate knowledge, typically provided in the vendor's documentation.

Optical and Laser Coordinate Surveying

Historically, coordinate surveys from optical instruments such as transits, theodolites, and electronic distance meters were the primary means of collecting geographic data. While these methods are slowly being replaced by satellite-based positioning and ground-based lasers, they are still quite common, and any competent GIS user should be familiar with optically based, field surveying methods. Spatial data layers are often produced directly from field surveys, or from field surveys combined with measurements on aerial photographs.

Surveying is particularly common for highly valued data. Real estate in upscale markets may be valued at hundreds to thousands of dollars per square meter. Zoning ordinances often specify the minimum distances between improvements and property boundaries. These factors justify precise and expensive coordinate surveys (Figure 5-18). Other commonly surveyed features are power lines, fiber optic cables, and utilities.

Plane surveying is horizontal surveying based on a planar (flat) surface. The flat surface assumption provides significant computational advantages, because the mathematics are substantially less complicated than those required for geodetic, or curved-Earth surveys. The flat surface in a plane survey is usually defined by a map projection, with a known point serving as the starting location for the survey. In U.S. urban areas, these are typically State Plane coordinates, or if defined, county coordinate systems.

In plane surveying, we typically assume *plumb* lines are perpendicular to the surface at all points in the survey. A plumb bob or weight is suspended from a string, and is assumed to hang in a vertical direction and intersect the plane surface at a 90° angle.

Figure 5-18: Surveying establishes the coordinates for most property lines. Field measurements of distance and direction are used to establish the set of vertices that define property boundary lines.This is the only way to collect these data, as the features are not visible on any other source.

Figure 5-19: A surveying instrument for collecting coordinate geometry data.

This is a valid assumption when the errors inherent with ignoring the Earth's curvature are small compared to the accuracy requirements of the survey, or to the errors inherent in the survey measurements themselves. The distance error due to assuming a flat rather than curved surface over 10 km (6 mi) is 0.72 cm (0.28 in). Therefore, plane surveys are typically restricted to distances under a few tens of kilometers. This restriction is met in many surveys, and a substantial majority of the lines and points surveyed to date have been measured using plane surveying methods. Plane surveying is sufficient for most subdivisions, public works, construction projects, and property surveys.

Historically, plane surveys have been conducted with optical instruments similar to those described for geodetic surveys. These instruments typically have angle gages in the horizontal and vertical planes and an optical sight, usually with some degree of telescopic magnification. The instruments have various names, including, in increasing order of sophistication and capabilities, a level, a transit, a theodolite, and a total station (Figure 5-19).

Plane surveys typically start at or are traceable back to a larger survey network through Bench Marks (precisely located points described in Chapter 3), or through local or project-specific control points established from high-accuracy GNSS positions. These marks or control points often serve as starting and ending points of a survey, to allow accuracy verification.

Distance and angle measurements are the primary field activities in plane surveying. Distances are measured between two *survey stations,* which are points occupied on the ground. The direction is specified by an angle between a standard direction, usually north or south, and the direction of the surveyed line between the two stations (Figure 5-20). The distance is in some standard units, for example, standard international meters.

There are two common ways of specifying angles. The first uses the azimuth. An azimuth angle is measured in a clockwise direction, typically relative to grid or geographic north (Figure 5-20). Azimuths vary from 0 to 360 degrees. Note that azimuths typically reference grid north, although they may also be specified relative to magnetic north, so care should be taken in clarifying which reference is used.

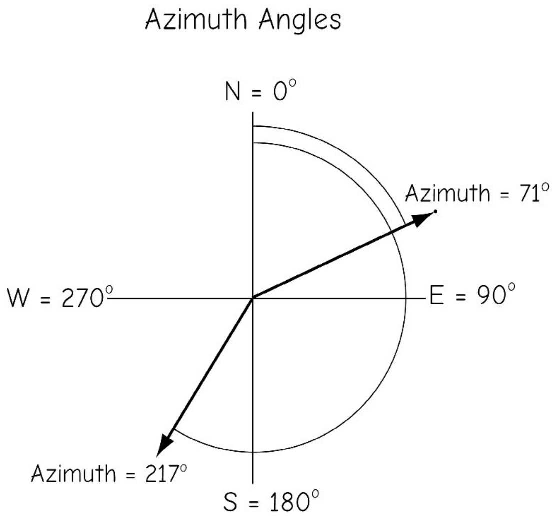


Figure 5-20: Angles in surveys may be reported as azimuth angles, measured clockwise relative to north, and ranging from 0 to 360.

Angles may also be specified by bearings, which use a north or south reference direction, an angle amount, and an east or west angle direction (Figure 5-21). The reference direction is either north or south, and the angle direction is east or west. The angle and direction are specified as “N71°E”, or “S37°W” .

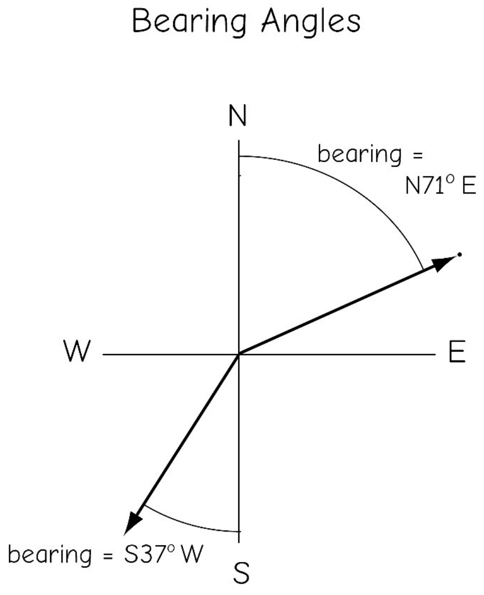


Figure 5-21: Directions may be specified as bearings, with a North or South reference, and an East or West angle.

We can convert between azimuth and bearing angles. Conversion from bearing to azimuth involves noting the reference direction (N or S), and adding or subtracting from constants, depending on quadrant (Figure 5-22). We often convert from bearings to azimuths when calculating positions from a sequential set of distance and angle measurements that form a survey.

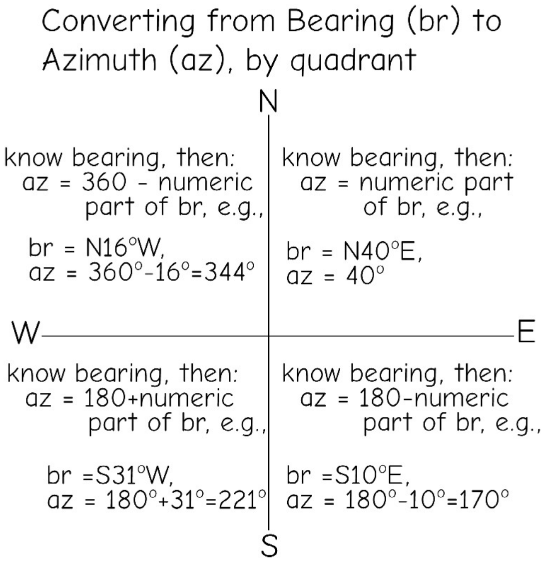


Figure 5-22: Conversion formulas from bearing to azimuth, by quadrant.

Many surveys are *traverses*, a series of connected lines that have a marked beginning and ending point. Traverses typically start at a known control point, or start at a point that has been referenced to a known control point. As described in the preceding sections, the control points are often part of a geodetic control network, or part of a subnetwork established by a municipal surveyor. A distance and angle are measured from the control point to the first survey station. *Coordinate geometry* (COGO) may be used to calculate the station coordinates. Subsequent distance and angle measurements may be taken, and in turn used to calculate the coordinates of subsequent stations. A traverse may be *open*, with a different beginning and ending point, or *closed*, with the traverse eventually connecting back to the starting location. Most of the millions of miles of property lines in North America have been established via plane surveys of open and closed traverses.

Coordinate geometry consists of a starting point (a station) and a list of directions (bearings) and distances to subsequent stations. The COGO defines a connected set of points from the starting station to each subsequent station. A sample COGO description follows:

"The starting point is a 1-inch iron rod that is approximately 102.4 ft north and 43.1 ft west of the northeast quarter of the southeast quarter section of section 16 of Township 24 North, Range 16 East, of the 2nd Principal Meridian. Starting from the said point, thence 102.7 ft on a bearing north 72.3 degrees east, to a 1-inch iron pipe; thence 429.6 ft on a bearing south, 64.3 degrees east to a 2-inch iron pipe..."

Basic trigonometric functions are used to calculate the coordinates for each survey station. These stations are located at the vertices that define lines or areas of interest. In the past, these distance and bearing data were manually plotted onto paper maps. Most survey data are now transferred directly to spatial data formats from the surveying instrument or associated software.

Field measurements may be directly entered and coordinate locations derived in the GIS software, or the coordinate calculations may be performed in the surveying instrument first. Many current surveying instruments contain an integrated computer and provide for digital data collection and storage. Coordinates may be tagged with attribute data in the field, at the measurement location. These data are then downloaded directly from a coordinate measuring device to a computer. Specialized surveying programs may be used for error checking and other processing. Many of these surveying packages will then output data in formats designed for import into various GIS software systems.

COGO calculations are illustrated on the left of Figure 5-23. Starting from a known coordinate, x_o, y_o, we measure a distance L and an angle θ. We may then calculate the distances in the x and y directions to another set of coordinates, x_1 and y_1. The coordinates of x_1 and y_1 are obtained by addition of the appropriate trigonometric functions. COGO calculations may then be repeated, using the x_1 and y_1 coordinates as the new starting location for calculating the position of the next traverse station.

The right side of Figure 5-23 shows a sequence of measurements for a traverse. Starting at x_s, y_s, the distance A and bearing angle, here 45°, are measured to station x_m, y_m. The bearing and distance are then measured to the next station, with coordinates x_n, y_n. Distances and angles are measured for all subsequent stations. Starting with the known coordinates at the starting station, x_s, y_s,

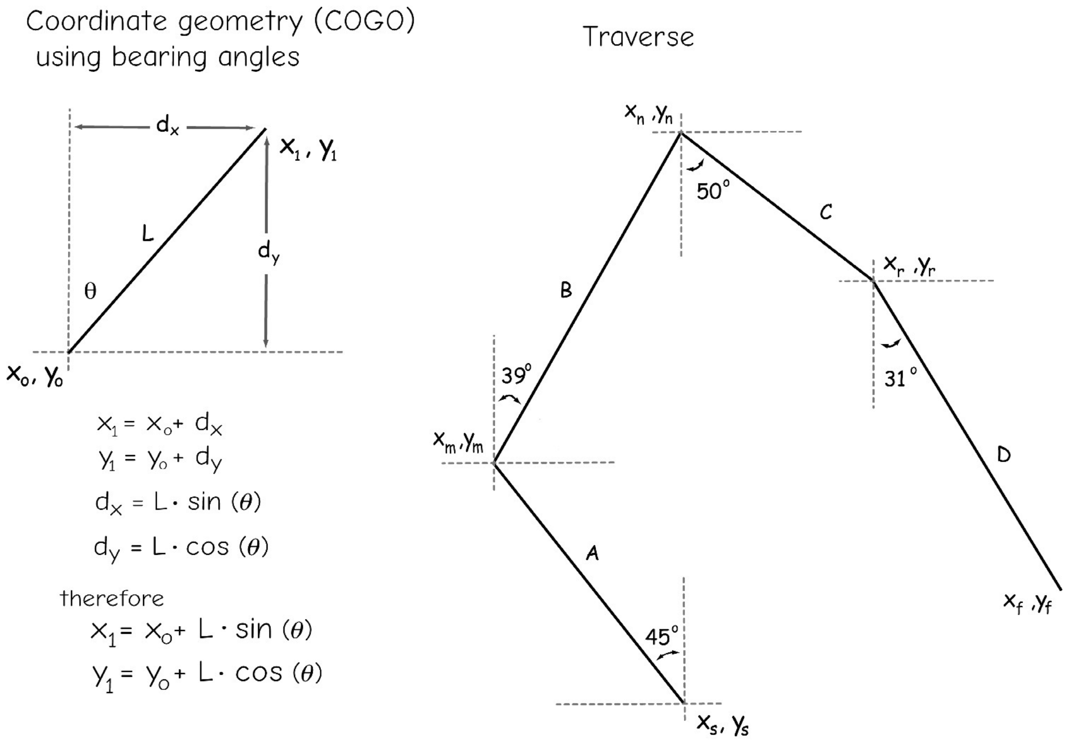


Figure 5-23: Coordinate geometry (COGO) allows the calculation of coordinate locations from open (shown above) or closed traverses. Distance and angle measurements are combined with trigonometric formulas to calculate coordinates.

coordinates for all other stations are calculated using COGO formulas.

Assigning the proper signs to d_x and d_y is important in COGO calculations. An incorrect sign for any leg of the traverse will propagate through all subsequent coordinates, causing each to be in error. The proper sign is obtained when directions are expressed as azimuths and a standard set of formulas is used.

The trigonometric sine and cosine functions return the proper magnitude and direction of d_x and d_y when applied to azimuth angles. If traverse angles are provided as bearings, they are typically converted to azimuths first, using the rules shown in Figure 5-22, remembering to convert the measured angles to radian units if those are required by the spreadsheet or computer language used for calculations. Sine and cosine values are then calculated and multiplied by the traverse leg distance, resulting in d_x and d_y values of the correct length and direction. Examples for all four quadrants are shown in Figure 5-24.

GNSS is now used for most measurements farther than a few hundred meters. COGO is more commonly applied in GIS when collecting data with a GNSS receiver in combination with a laser rangefinder. Laser rangefinders emit a focused, coherent beam of light to calculate distance. The maximum range depends on the size and reflective properties of the target, but many moderately priced lasers are accurate up to several hundred meters. Electronic compasses must be periodically calibrated and adjusted for proper magnetic declination, but can be quite accurate. Rangefinders often also have vertical angle gages, because all measurements are assumed parallel to the datum plane, and non-horizontal measurements must be adjusted. Careful laser rangefinders measurements can be accurate to a centimeter over a 100 m distance, improving the efficiency of GNSS data collection.

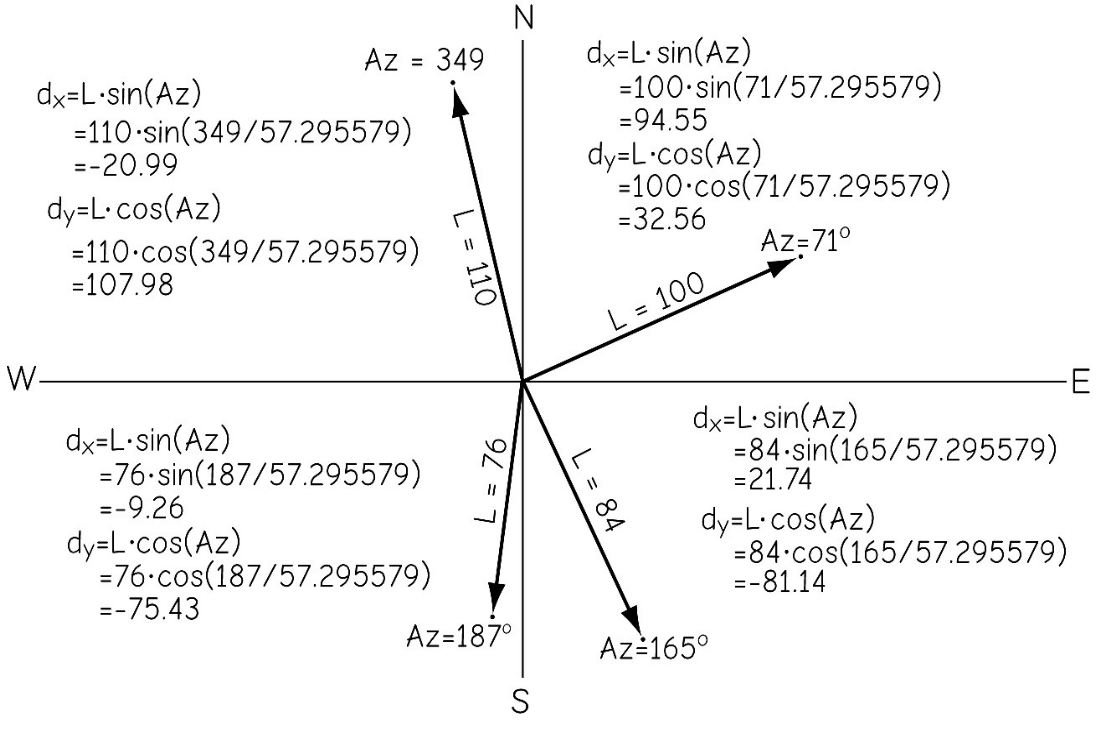


Figure 5-24: Example calculations of dx and dy values, given a survey distance (L) and an azimuth angle (Az). Note that the sine and cosine functions return the correct sign as well as magnitude, so that the dx and dy values are positive or negative, as appropriate. Also note that many spreadsheets and coding applicaitons default to radian units in sine and cosine functions, hence the division of degree angles by the conversion factor 57.295579.

Figure 5-25:Three-dimensional scanning lasers use coordinate geometry to record comprehensive x, y, and z coordinate data for GIS. Here, a scanning laser is shown (foreground) with a portion of the 3-D laser measurements (left) superimposed on the bridge (courtesy Leica Geosystems).

Terrestrial, three-dimensional lasers are rapidly becoming common, and although currently used primarily for structure analysis, they may be used for GIS data entry. These systems emit narrow, directed laser pulses. By carefully measuring the horizontal and vertical angles relative to the established coordinate system, these laser distance measurements can be converted to three-dimensional coordinates via coordinate geometry. GNSS systems typically provide the location of the laser at the time of data capture, but additional measurements are often necessary to establish an initial or reference pointing direction. These data are then used to generate two- or three-dimensional data layers for spatial data bases.

Terrestrial three-dimensional laser systems collect billions of points, and collections from multiple locations must be combined through three-dimensional reconstructive models to create complete digital representations of real-world objects (Figure 5-25). As software and computer systems improve, three-dimensional terrestrial lasers will become common.

GNSS Applications

Tracking, navigation, field digitizing, and surveying are the main applications of GNSS. Navigation is finding a way or route, and tracking involves noting the location of objects through time. A common example is tracking delivery vehicles in near real time. Large delivery and distribution organizations frequently require information on the location of a fleet of vehicles. Vehicles equipped with a GNSS receiver and a radio or phone link may report back to a dispatch office every few seconds. In effect, the dispatcher may have a real-time map of the vehicle location. Icons on a digital map are used to represent vehicles, and a quick glance can reveal which vehicle is nearest a delivery or retrieval site, or which driver overly frequents a donut shop.

Navigation is a second common GNSS application. GNSS receivers have been developed specifically for navigation, with digital maps or compasses set into on-screen displays (Figure 5-26). These GNSS receivers and digital maps are extremely specialized GIS systems. These systems are useful when collecting or verifying spatial data, such as to navigate to the approximate vicinity of field measurement plots.

Field Digitization

Field digitization is a primary application of GNSS in GIS. Data may be recorded directly in the field to update point, line, or area locations. Features are visited or traversed in the field, and an appropriate number of GNSS fixes collected. GNSS receivers have been carried in automobiles, on boats, bicycles, and helmets, or by hand to capture the coordinate locations of points and boundaries (Figure 5-27).

GNSS data are often more accurate than data collected from the highest-quality digital images. For example, RTK GNSS data typically have accuracies better than 5 cm, and often below 2 cm, while accuracies are often near 15 to 50 cm for the highest-resolution satellite images, and for national aerial image programs. Precise differential correction of carrier phase GNSS data often yield centimeter-level accuracies, far better than can be obtained from digitizing almost all images.

GNSS is often used to directly digitize new control points. Remember that control points are used to correct and transform image data or maps to real-world coordinates. Aerial images may be available, and

Figure 5-26: A GPS receiver developed for marine (left) and aerial navigation (right, courtesy Garmin Corp. and Trimble Ltd.).

Figure 5-27: Line features may be field digitized via GPS, as in this example of a GIS/GPS system mounted in a tractor. Data display and digitizing software are used to record coordinates collected by a GPS receiver (above). An antenna placed on a pole or rack (right) reduces obstruction (courtesy Jake Leguee, left, and G. Johnson, Ducks Unlimited, right).

the coordinates may be unknown for features visible on the aerial image. Control points may be difficult or impossible to obtain directly from surveys or from the information plotted on existing maps, particularly when graticule or gridlines are absent. GNSS offers a direct method for measuring the coordinates for potential control points represented on the image or map. Road intersections or other points may be identified and then visited with a GNSS receiver.

GNSS-measured control points are the basis for almost all current projects that perform analytical correction of aerial imagery (see Chapter 6). Most image data are not initially in a map coordinate system, yet images are often particularly useful for developing or updating spatial data. Aerial photographs contain detailed information. However, aerial photographs are subject to geometric distortion. These errors may be analytically corrected through suitable methods (see Chapter 6), but these methods require several control points per image, or at least per project, when multiple, overlapping aerial photographs are used. GNSS significantly reduces the cost of control point collection, thereby making single- or multiphoto correction a viable alternative for most organizations that collect spatial data.

Feature digitization with GNSS often involves the capture of both coordinate and attribute data. Typically, the GNSS receiver is activated and detects signals from a set of satellites. A file is opened and position fixes are logged at a set rate, such as every two seconds. Attribute data may also be entered, either while the position fixes are being collected, or before or after positional data collection. In some software, the position fixes may be tagged or identified. For example, a specific corner may be tagged while digitizing a line. Multiple features may be collected in one file and the identities maintained via attached attributes. Data are processed as needed to improve accuracy, and converted to a format compatible with the GIS system in use. GNSS data collection and data reduction tools often provide the ability to edit, split, or aggregate collected data, for example, converting multiple fixes into a single point average. These functions may be applied for all position fixes in a file, or for a subset of position fixes embedded in a GNSS file.

Large, field portable displays and advanced editing software may be combined with real-time differential correction to improve field digitizing. Tablet computers are available with large color screens (Figure 5-28). Scanned digital images may be dis-

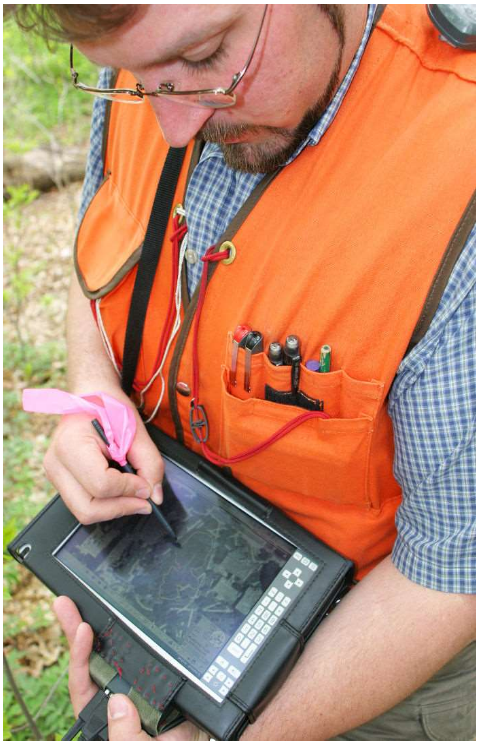


Figure 5-28: Ruggedized tablet computers may be carried in the field and used for data entry. Large screens allow efficient display and field editing of spatial data.

played with existing digital data. New data may be input via a GNSS receiver, in real time, or via penstrokes on the screen. The operator may digitize new features, edit old ones, or perform some combination of the two while in the field (Figure 5-29). Snapping tolerances, maximum overshoots, and all other digitizing controls may be applied in the field, much like when digitizing on-screen in an office.

GNSS field digitization is most commonly used for collection of point and line features. Multiple position fixes provide higher accuracies and are often collected for point locations, and for important vertices in line data. However, GNSS data collection for line and area features suffer from a number of unique difficulties. First, it takes considerable time to traverse an area, so relatively large parcels or many small parcels may be impractical to digitize in the field. Second, multiple representations of the same boundary may occurs when digitizing polygonal features. Attempting to retrace the common boundary wastes time and provides redundant and conflicting data while field digitizing. The alternative is to digitize only the new lines, and snap to "field-nodes," much as when capturing data using a coordinate digitizer (see Chapter 4). This

Figure 5-29: Features may be entered and edited in the field using a GPS receiver and appropriate software (courtesy Trimble, Ltd.).

method is often used, with subsequent editing in a GIS.

GNSS field software are often optimized to streamline the input of attributes that are associated with spatial data. Forms may provide menus, pick lists, and variable entry boxes in a predetermined order. These software often improve attribute data accuracy, in part by helping avoid blunders. For example, the entry options for a specific attribute such as fire hydrant color may be restricted to red, green, or yellow from a "pick list," if those are the only possible values. These attribute entry forms also increase completeness, in part by ensuring that every variable is presented to the operator, and these forms may also be configured to show a warning when all variables have not been entered.

Single Fix vs Averaged Accuracy

Lines or polygons digitized with GNSS receivers usually have larger errors than those reported in the technical specifications or marketing literature. Figure 5-30 illustrates this. A stream network was digitized with a professional grade GNSS receiver, with an advertised 50 cm (20 inch) "average" error. The unit was configured to maximize single-fix accuracy. Multiple passes up and down the streams were digitized, at normal walking speeds. The digitized tracks (thinner, jagged lines) vary notably about the true stream location (thick, smoother lines) with errors often in the 3 to 5 meter range (10 to 16 feet), and as large as 25 meters (82 feet). These observed errors are much larger than the reported average values for at least two reasons.

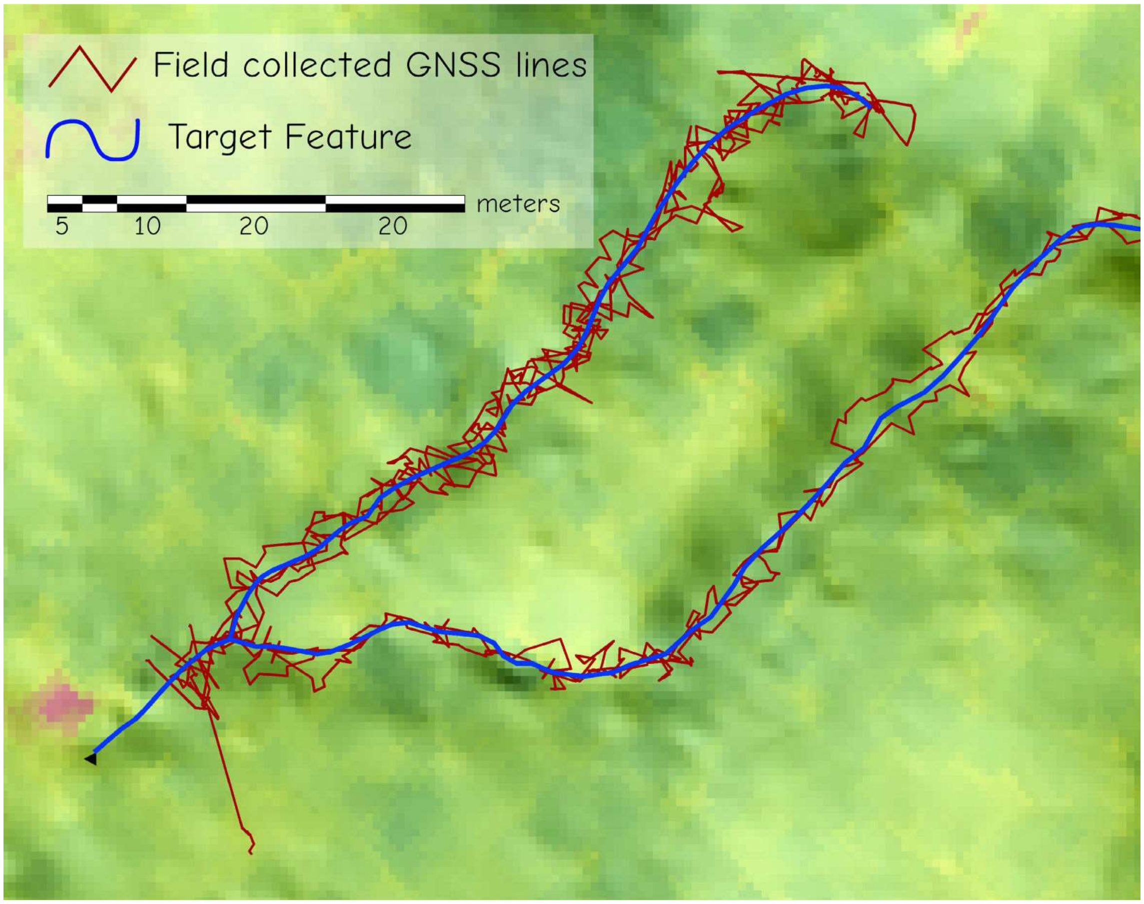


Figure 5-30: An example of several field-digitized lines, using a professional-grade GNSS receiver in a steep ravine with dense forest cover. Single fixes vary substantially about the true locations. The single fix errors are larger than averages and other summary statistics reported for a receiver, caused by terrain and vegetation obstruction, reducing PDOPs and increasing multi-path signals.

First, manufacturers usually report an average or distributional accuracy, with some description of the statistics used. Often they use terms like CEP, the circular error probability, or one-sigma (1σ) errors, measures of how frequently you'd expect to get an error larger than the reported size. These averages give you an optimistic view of single-fix accuracy, because averaging reduces variability, and so the mean error tends toward a stable, lower value. Errors to the north cancel out errors to the south, and so a mean of even a few fixes is much better behaved than a single fix. One-sigma or CEP measures are a bit more helpful, but again, they usually don't tell us much about the frequency of large errors. Unfortunately, we typically digitize single fixes when collecting lines or polygon boundaries, so the differences between single fix and average errors are important.

Second, accuracies are often specified under ideal conditions, in unobstructed locations with no trees, buildings, or mountains to block or reflect satellite signals. While these conditions reign for much of the world, for many places they do not, decreasing single-fix accuracy. The data in Figure 5-30 were collected under dense forest, so show substantial scatter due to lower PDOPs and higher multi-path common in obstructed environments. As described in the next section, we typically use several strategies during field data collection to increase GNSS accuracy and reduce the need for post-collection editing. But even employing these, we often must manually edit line data or polygon boundaries when they are field digitized, to remove the occasional large error.

Field Digitizing Accuracy and Efficiency

Field GNSS collections are affected by an obstructed sky. Terrain, trees, buildings, or other objects block out portions of the sky, causing temporary interruptions in satellite reception, or forcing the GNSS receiver to estimate position from a constantly changing set of satellites. Maximum collection rates while digitizing in the field with a GNSS receiver are typically near one fix per second. Obstructions may increase collection intervals between fixes to several seconds or minutes.

Obstructions may halt GNSS field digitization entirely if they reduce the number of visible satellites to three or fewer. Sky obstructions reduce the efficiency of field digitization because more time is spent collecting a given number of fixes, and personnel must wait for the satellite constellation to change when satellites are too few or poorly distributed. Alternately, they may collect fewer positions, thereby reducing positional accuracy.

Reductions in the efficiency of GNSS digitization depend on the nature of the obstruction, the type of equipment, the equipment configuration, and satellite number and position. GNSS signals may pass through foliage when collecting data below a forest canopy, although signals become weaker as they pass through several canopy layers. Satellite signals are blocked by stems and branches, though individual satellites are typically obstructed by stems for relatively short durations. Under dense canopy, the available satellite constellation may change frequently; slightly changing the position of the GNSS antenna, by raising or lowering it, may result in a new constellation of visible satellites. Despite these efforts, efficiency reductions may be substantial, doubling or tripling collection times, but single-fix collection times rarely take longer than a few seconds to minutes when using modern, professional-grade receivers and when forest canopy is the primary sky obstruction. Collection times will increase

correspondingly when multiple fixes are required per feature.

Terrain may block satellites, a significant problem when the blocked satellites are greater than 15° above the local horizontal plane. Satellites less than 15° above the local horizontal plane are of limited use, even in open conditions, because they exhibit large range errors due to atmospheric interference. GNSS receivers designed for GIS data collection typically provide settings that automatically reject satellites below a specified horizon angle.

Terrain obstructions often rise above 15°, such as when mountains, hillslopes, buildings, or canyon walls reduce the number of visible GNSS satellites (Figure 5-31). Terrain obstruction often reduces collection efficiencies and accuracies. Because the GNSS signals do not pass through soil, rock, wood, or concrete, any obstructed satellite cannot be used for GNSS positioning. In some instances, a short wait may result in a rearrangement of the satellite constellation, such as from point c to point b in Figure 5-31. However, on average, an obstructed sky results in a reduced constellation of GNSS satellites and higher PDOPs when compared to flat terrain. This problem is particularly vexing in urban settings because the horizon angles change substantially over short distances. This makes it difficult to predict when GNSS satellite coverage will be adequate, and thus plan data collection efforts.

Forest, terrain, and building effects may occur together, further reducing accuracies and decreasing efficiencies. This is a common occurrence in forested, mountainous terrain, and in urban areas with both tall buildings and mature trees.

The use of a *range pole* is perhaps the easiest, most common, and often most effective method to improve collection efficiency (Figure 5-32). A range pole is an extendable pole on which a GNSS antenna is mounted. A range pole is often particularly effective in urban and forested conditions, where canopy gaps and building obstructions vary vertically. A range pole facilitates the search for an acceptable set of satellites. The antenna is raised and lowered during data collection as the satellite constellation changes through time and long pauses are encountered. A range pole is perhaps most useful when digitizing point features or important vertices in

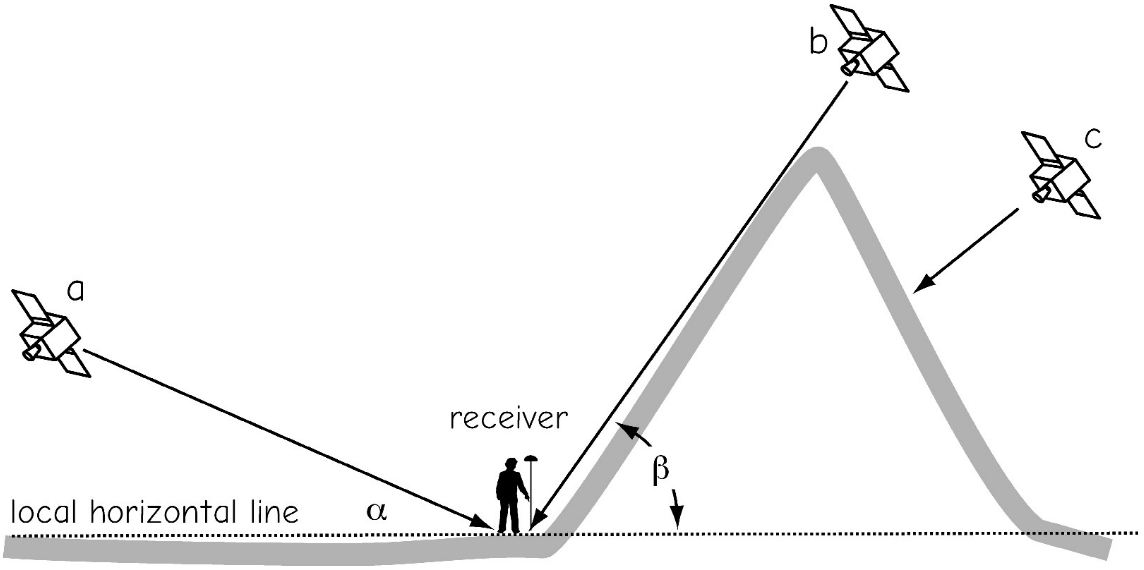


Figure 5-31: GPS satellite signals may be blocked by terrain or built structures. This reduces the constellation of available satellites, increasing error and reducing collection efficiency. Here, satellites a and b are visible with corresponding local horizon angles of α and β. The signal from satellite c is blocked by local terrain.

Figure 5-32: A range pole in use with a GNSS receiver.

line features, when the receiver remains stationary.

Handheld or backpack-mounted poles commonly improve efficiency when digitizating with GNSS. Raising the antenna just a few meters off the ground, avoids low obstructions such as the body and thick skull of the human operator.

There is often a trade-off between accuracy and efficiency during field digitization, particularly in obstructed locations (Figure 5-33). This series of graphs shows data collected by Scrinzi and Floris (1998), in rough terrain and under forest canopy. These results are dated and so the absolute numbers are probably pessimistic for newer equipment, but the general patterns still hold true. As shown in the leftmost graph of Figure 5-33, they found that 100% of the possible fixes may be collected when the average horizon angle is near 15°. They also collected data at various points in hilly terrain, where the horizon angle was greater because mountains and ridges block lower portions of the sky. Efficiency dropped to near 70% as average horizon angle increased to near 30° . Collections took about 30% longer or fixes were 30% less frequent when in valley locations compared to flat terrain. However, the leftmost graph gives optimistic estimates in that it shows efficiencies when accepting fixes with any PDOP. Since we know accuracy decreases at higher PDOPs, we often set a maximum PDOP threshold. This increases the accuracy and increases collection times when using a GNSS.

The center and right graphs in Figure 5-33 show the increased impact of horizon angle when keeping PDOPs below specified thresholds. The center graph shows that collection efficiency falls off more rapidly when the PDOP threshold is set at 8. Collection efficiency is approximately 50% when horizon angles average 30° — in other words, it takes approximately twice as long to collect the same amount of data, or approximately one-half the fixes are collected in the same amount of time. These effects are magnified when PDOPs are restricted to less than 4 (Figure 5-33, right graph). Approximately 20% of the possible position fixes were recorded at a 30° average terrain angle and PDOP threshold of 4, suggesting only one in five position fixes will be obtained. While Figure 5-33 was generated with a specific, high-quality receiver optimized for field digitization, the general patterns are true for all currently available GNSS systems — efficiency decreases in obstructed terrain, and the rate of decrease changes with the allowable PDOP. As GNSS receivers improve, and can measure GPS, GLONASS, Galileo, and Compass GNSS simultaneously, efficiencies and accuracies in obstructed environments have substantially increased, such that reasonable effi-

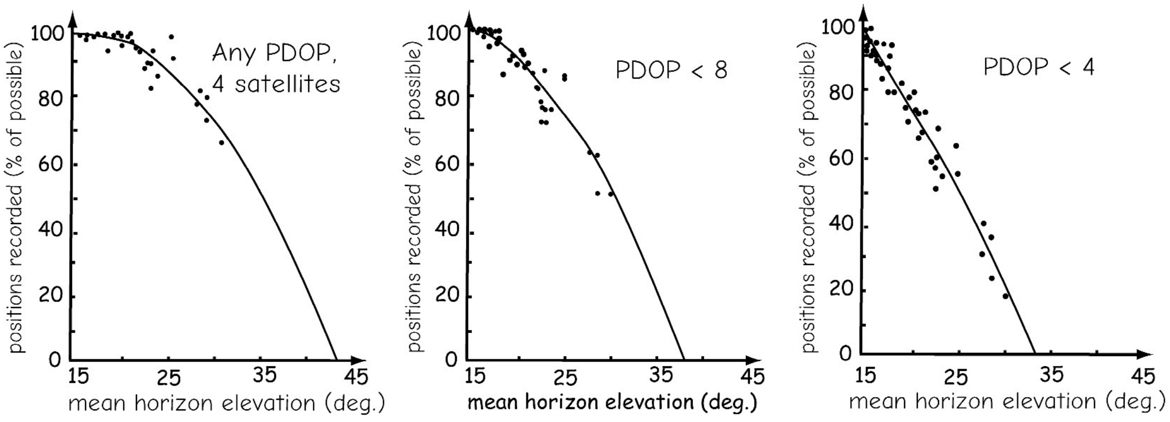


Figure 5-33:The percentage of GPS position fixes that are successfully collected decreases in valleys, or in any other location where the angle to the horizon increases. This may be offset somewhat by allowing poorer (larger) PDOPs, as shown in the leftmost vs. rightmost figures above (adapted from Scrinzi and Floris, 1998).

ciencies are currently obtained under most conditions.

We may improve the efficiency of GNSS digitization by altering PDOP and signal strength thresholds, but this often comes at the expense of decreased accuracy. Sophisticated receivers allow multiple settings; a target PDOP, above which the receiver will search for better satellite constellations, and a maximum PDOP, above which data collection will cease. This allows the user to balance the trade-off between accuracy and efficiency.

Some GNSS receivers allow adjustments in the threshold for acceptable signal strength. For example, satellite signals that pass through a forest canopy are weaker. Including these weaker signals improves the number and often the distribution of satellites, thereby increasing collection efficiency and perhaps accuracy. However, weak signals may also result from reflected or multipath transmissions. As described earlier, multipath signals have larger range errors. Lowering the threshold for acceptable signal strength is likely to increase positional error, as it increases the likelihood of multipath measurements. However, some data are often better than none, and lowering the PDOP threshold for collection is sometimes the only way to collect data.

GNSS receivers specifically designed for GIS data collection may be fitted with a sensitive antenna that also reduces multipath reception. Manufacturers have invested substantially in optimizing antenna design and collection systems to control these multiple trade-offs. The availability of specifically optimized antennas is a primary difference between GIS-grade receivers and recreational receivers costing much less. Recreational receivers are substantially less accurate in obstructed terrain. Recreational receiver thresholds for signal strength or PDOP are often configured for highest efficiency and thus lowest accuracy under obstructed conditions, and these thresholds often may not be adjusted by the user. Irrespective of the equipment, there is a trade-off between the acceptable signal strength and the introduction of multipath errors. Setting the maximum acceptable PDOP higher or acceptable signal strength thresholds lower will increase efficiency of collection, but often at the cost of increased error.

Rangefinder Integration

There are other limits to GNSS data collection. For example, the need to occupy every vertex and node in the field is a primary drawback of GNSS digitizing. Sometimes, it may be dangerous to physically place the GNSS receiver over each point, for example, when a stream to be digitized is in a field full of rutting buffalo. Features may be difficult to reach, costing the user more time in travel than in GNSS data collection. This is particularly common when point features to be digitized are widely dispersed. Features may be numerous, intervisible, but separated by a barrier, for example, a sequence of fence posts or power poles on opposite sides of a limited access highway.

Figure 5-34: A laser rangefinder may substantially improve the efficiency of field data collection with GPS. Here, a system integrates a binocular unit to automatically calculate positions from GPS measurements and distance and angle observations. Also note the range pole to raise the GNSS antenna above the collector's head (courtesy Leica Geosystems).

Peripheral measuring devices, such as laser rangefinders, may be attached to GNSS data collectors to substantially improve field data collection (Figure 5-34). These devices typically measure distance with a laser and direction with a compass. Measurements are made from each occupied GNSS point to the nearby features of interest. The target coordinate calculations are often automatic because direction is measured with an integrated electronic compass. The rangefinder is pointed at the feature to be digitized. The system calculates the observer's position from the GNSS, and this position is combined with distance and angle measurements in coordinate geometry to calculate the feature coordinates. The person operating the GNSS/laser rangefinder may stand in one location and collect positions for several to tens of features, thereby saving substantial travel time. These systems are most often used to inventory point features such as utility poles, signs, wells, trees, or buildings.

Laser rangefinders are available that can measure features at distances up to 600 m (2000 ft). Realized accuracies depend on both the quality of the GNSS receiver and the distance measuring subsystem. However, submeter accuracies are possible under open sky conditions.

GNSS Height Measurement

GNSS are often the easiest way to measure new heights, but care must be taken to ensure heights are relative to an appropriate vertical datum. GNSS heights are typically determined as a height above an ellipsoid, or HAE, and many less expensive GPS receivers have no options to report any other height. The user should carefully read the documentation to determine the type of height reported by the GNSS unit. Typically, it is ellipsoidal height, although some units provide an estimate of orthometric height. Since the two can differ by up to 100 meters (330 feet), the user should determine which height is reported.

As described in Chapter 3, our standard height reference is a vertical datum and not an ellipsoid, so we must convert any provided HAEs to an orthometric height, relative to a datum, prior to most uses. As explained in Chapter 3, we calculate the orthometric height via the equation:

$$H = h - N \tag{5.2}$$

The GNSS often provides h, the ellipsoidal height, and we may use spatial models developed by most governments to estimate N, the geoidal height, for any location. In the U.S., these models have been developed and documented by the National Geodetic Survey, and are available at a geoids page (http://www.ngs.noaa.gov/GEOID/). These models have been incorporated into most GNSS receivers designed for GIS data collection, and so the conversion may be transparent, or available as a system setting.

If the receiver does not support geoidal height estimation, and if the user has no access to computing or software required for the NGS geoid models, one may estimate geoidal height by referencing a nearby control point, for example, an NGS control data sheet in the United States. These sheets, also described in Chapter 3, usually provide geoidal heights for listed points. Geoidal heights

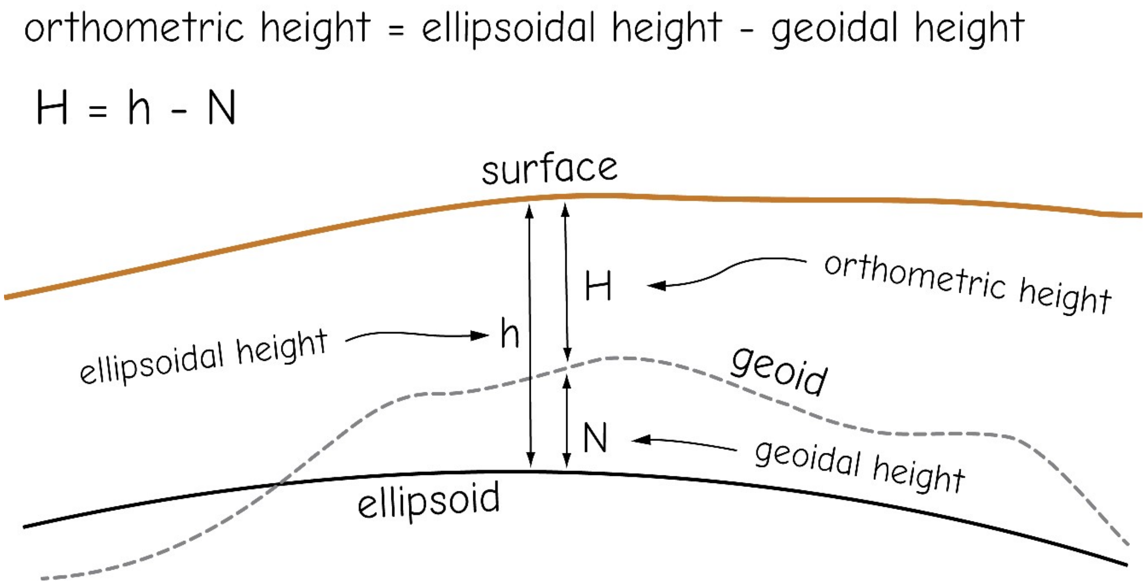


Figure 5-35: Calculation of orthometric height from ellipsoidal height. Most GNSS systems report ellipsoidal heights, which must be converted to orthometric heights before most uses. We may estimate geoidal height in several ways, including developed geoidal models, or nearby geodetic control points.

do not vary rapidly across space for most regions of the globe, and the spatial variability of geoidal heights may be estimated by retrieving heights for several nearby NGS control points. The nearest, an average, or some similar combination of geoidal heights may be used in equation (5.2) to calculate orthometric height, given a measurement of orthometric height from the GNSS.

GNSS Tracking

GNSS tracking of people, vehicles, packages, or animals is an innovative and growing application of GNSS. GNSS receivers are routinely placed on trucks, ships, buses, boats, or other transport vehicles. These receivers are often part of systems that include information on local conditions, speed of travel, and perhaps the condition of the shipped equipment or cargo.

GNSS is also increasingly applied to track individual organisms. This is revolutionizing animal movement analysis because of the frequency and density of points that may be collected (Figure 5-36). More position fixes can be collected in a month using GNSS equipment than may be collected in a decade using alternative methods.

Animal movement analysis has long been based on observation of recognizable individuals. Each time a known animal is seen, the location is noted. The number of position fixes is often low, however, because some animals are difficult to spot, elusive, or live in areas of dense vegetation or varied terrain. Early alternatives to direct human observations were based primarily on *radio-*

Figure 5-36: A wildebeest fit with a GNSS tracking collar. The antenna is visible as the white patch on top of the collar, and the power supply and data logging housing is visible at the bottom of the collar. Animal position is tracked day and night, yielding substantially improved information on animal activity and habitat use. (Courtesy Gordon T. Carl, Lotek Wireless Inc.)

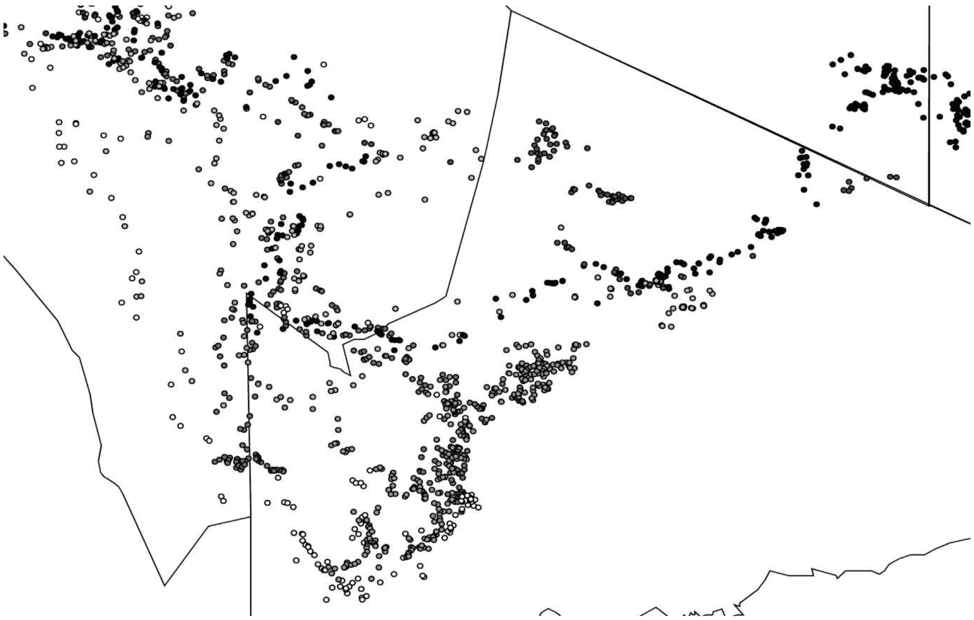


Figure 5-37: GNSS tracking data for wildebeest in the Serengeti and Ngorongoro Crater regions of Tanzania. Studies of travel routes and habitat use by migratory animals are substantially improved by GPS data collection (courtesy S. Thirgood, A. Mosser, and M. Borner).

telemetry. Radiotelemetry involves the use of a transmitting and receiving radio unit to determine animal location. A transmitting radio is attached to an animal, and a technician in the field uses a radio receiver to determine the position of the animal. Measurements from several directions are combined, and the approximate location of the animal may be plotted.

GNSS animal tracking is a substantial improvement over previous methods. GNSS units are fit to animals, usually by a harness or collars (Figure 5-36). The animals are released, and positional information recorded by the GNSS receiver. Logging intervals are variable, from every few minutes to every few days, and data may be periodically downloaded via a radio link. Systems may be set up with an automatic or radio-activated drop mechanism so that data may be downloaded and the receiver reused. While only recently developed, GNSS-based animal tracking units are currently in use on all continents in the study of threatened, endangered, or important species.

GNSS tracking for individual or fleets of vehicles typically involves a number of

subsystems (Figure 5-38). GNSS receivers and radio transmitters must be placed on each vehicle to record and transmit position. Satellite or ground-based receiver networks collect and transmit positional and other data to a computer running a tracking and management program that may be used to display, analyze, and control vehicle movement. Information or instructions may be passed back to vehicles on the road.

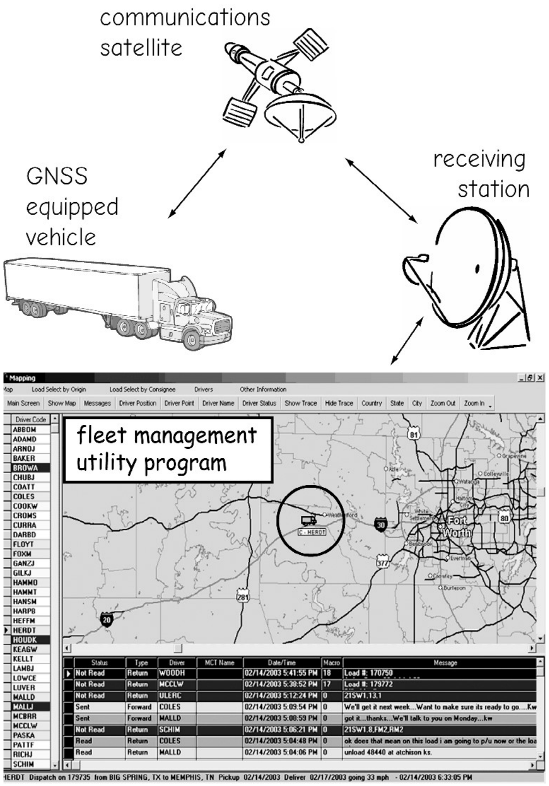


Figure 5-38: Real-time tracking via GPS substantially improves vehicle fleet management, particularly when combined with other data in a GIS.

GNSS-aided vehicle management may be combined with other spatial data in a GIS framework to add immense value to spatial analyses. Vehicle location can be monitored in real time, and compared to delivery locations. Delivery planning may be optimized and delivery windows specified with much greater accuracies. This in turn may substantially reduce costs, increase data gathering, and improve profits for participating businesses. Transport may be dispatched more efficiently, recurring problems analyzed, and solutions more effectively tailored.

Imaginative uses of GNSS are arising almost daily as this technology revolutionizes positional data collection. GNSS equipment has been interfaced with grain harvesting equipment. Grain production is recorded during harvest, so that yield and grain quality are mapped every few meters in a farm field. This allows the farmer to analyze and improve production on a site-specific basis, for example, by tailoring fertilizing applications for each square meter in the field. The mix of fertilizers may change with position, again controlled by a GNSS receiver and software carried aboard a tractor.

Summary

GNSS is a satellite-based positioning system. It is composed of user, control, and satellite segments, and allows precise position location quickly and with high accuracy.

GNSS is based on range measurements. These range measurements are derived from measurements of a broadcast signal that may be either coded or uncoded. Uncoded, carrier phase signals are the basis for the most precise position determination, but are of limited use for locating features due to measurement requirements. Code phase measurements are primarily used for feature collection and entry into GIS. Range measurements from multiple satellites may be combined to estimate position.

GNSS positional estimates contain error due to uncertainties in satellite position, atmospheric and ionospheric interference, multipath reflectance, and poor satellite geometry. These uncertainties vary in time and space.

There are a number of ways to ensure the highest accuracy when collecting GNSS data. Perhaps the greatest improvement comes from differentially correcting GNSS positions. Differential correction is based on simultaneous GNSS measurements at a known base location and at unknown field locations. Errors are calculated for each position fix at the base station, and subtracted to the field collections to improve accuracy. Accuracy may also be improved by collecting with low PDOPs, averaging multiple position fixes for each feature, avoiding multipath or low horizon signals, and using a GNSS receiver optimized for accurate GIS data collection.

GNSS is most commonly used in GIS to digitize features in the field, either for primary data collection, to update existing data, or for secondary data collection, to support orthoimage creation. Terrain, buildings, or tree canopy commonly obstruct the sky, leading to reduced accuracy and efficiencies. Modifying PDOP and signal strength thresholds to account for these obstructions may increase collection efficiencies, but often at the expense of reducing accuracies. Specialized antennas and firmware help, and these are commonly available on GIS-grade receivers, but not on commercial receivers.

GNSS receivers are also used for tracking, navigation, and field surveying. Vehicle tracking applications require GNSS, transmission, and interpretation subsystems, and are becoming widely applied. Animal and human movements are increasingly being tracked via GNSS.

Suggested Reading

Abidin, H. (2002). Fundamentals of GPS signals and data. In Bossler, J. (Ed.). *Manual of Geospatial Science and Technology*. London: Taylor and Francis.

Awange, J.L. (2012). *Environmental Monitoring Using GIS*. Berlin: Springer-Verlag.

Bobbe, T. (1992). Real-time differential GPS for aerial surveying and remote sensing. *GPS World*, 4:18–22.

Deckert, C.J., Bolstad, P.V. (1996). Forest canopy, terrain, and distance effects on global positioning system point accuracy. *Forest Science*, 62:317-321.

Dominy, N.J., Duncan, B. (2002). GPS and GIS in an African rain forest: Applications to tropical ecology and conservation. *Conservation Ecology*, 5:537-549.

Dow, J.M., Neilan, R.E., Rizos, C. (2009). The International GNSS Service in a changing landscape of Global Navigation Satellite Systems. *Journal of Geodesy* 83:191-198.

Dwolatzky, B., Trengove, E., Struthers, H., McIntyre, J.A., Martinson, N.A. (2006). Linking the global positioning system (GPS) to a personal digital assistant (PDA) to support tuberculosis control in South Africa: a pilot study. *International Journal of Health Geographics*, 5:34.

Fix, R.A. Burt, T.P. (1995). Global Positioning Systems: an effective way to map a small area or catchment. *Earth Surface Processes and Landforms*, 20:817-827.

Gao, J., Liu, Y.S. (2001). Applications of remote sensing, GIS and GPS in glaciology: a review. *Progress in Physical Geography*, 25:520-540.

Gao, J. (2002). Integration of GPS with remote sensing and GIS: Reality and prospect. *Photogrammetric Engineering & Remote Sensing*, 68:447-453.

Haibo, H., Jinlong, L., Xu, J., Guo, H., Wang, A. (2014). Performance assessment of single- and dual-frequency BeiDou/GPS single-epoch kinematic positioning. *GPS Solutions*, 18:393-403.

Jagadeesh, G.R., Srikanthan, T., Zhang, X.D. (2004). A map matching method for GPS based real-time vehicle location. *Journal of Navigation*, 57:429-440.

Johnson, C.E., Barton, C.C. (2004). Where in the world are my field plots? Using GPS effectively in environmental field studies. *Frontiers in Ecology & the Environment*, 2:475-482.

Kaplan, E.D., Hegarty, C. J.(2006). *Understanding GPS: Principles and Applications*. Norwood: Artech House.

Kennedy, M. (1996). *The Global Positioning System and GIS*. Ann Arbor: Ann Arbor Press.

Mintsis, G., Basbas, S., Papaioannou, P., Taxiltaris, C., Tziavos, I.N. (2004). Applications of GPS technology in the land transportation system. *European Journal of Operational Research*, 152:399-409.

Næsset, E., Jonmeister, T. (2002). Assessing point accuracy of DGPS under forest canopy before data acquisition, in the field, and after processing. *Scandinavian Journal of Forest Research*, 17:351-358.

Odolinski, R., Teunissen, P.J. (2016). Single-frequency, dual-GNSS versus dual-frequency, single-GNSS: a low-cost and high-grade receivers GPS-BDS RTK analysis. *Journal of Geodesy*, 90:1255-1278.

Scrinzi, G., Floris, A. (1998). Global Positioning Systems (GPS), una nuova realtà nel rilevamento forestale, Atti del Convegno "Nuovi orizzonti per l'assestamento forestale" 14-56.

Small, E.D., Wilson, J.S., Kimball, A.J. (2007). Methodology for the re-location of permanent plot markers using spatial analysis. *Northern Journal of Applied Forestry*, 24:30-36.

Toledo-Moreo, R., Betaille, D., Peyret, F., Laneurit, J. (2009). *IEEE Journal of Selected Topics in Signal Processing*, 3:798-809.

Thirgood, S., Mosser, A., Tham, S., Hopcraft, G., Mwangomo, E., Mlengeya, T., Kilewo, M., Fryxell, J., Sinclair, A.R.E., Borner, M. (2004). Can parks protect migratory ungulates? The case of the Serengeti wildebeest. *Animal Conservation*, 7:113-120.

Van Sickle, C. (2008) *GPS for Land Surveyors, 3rd edition*. Boca Raton: CRC Press.

Welch, R., Remillard, M., Alberts, J. (1992). Integration of GPS, remote sensing, and GIS techniques for coastal resource management. *Photogrammetric Engineering and Remote Sensing*, 58:1571-1578.

Wilson, J.P., Spangrud, D.S., Nielsen, G.A., Jacobsen, J.S., Tyler, D.A. (1998). GPS sampling intensity and pattern effects on computed terrain attributes. *Soil Science Society of America Journal*, 62:1410-1417.

Study Questions

5.1 - Describe the general components of GNSS, including the three common segments and what they do.

5.2 - What is the basic principle behind GNSS positioning? What is a range measurement, and how does it help you locate yourself?

5.3 - Describe the GNSS signals that are broadcast, and the basic difference between carrier and coded signals.

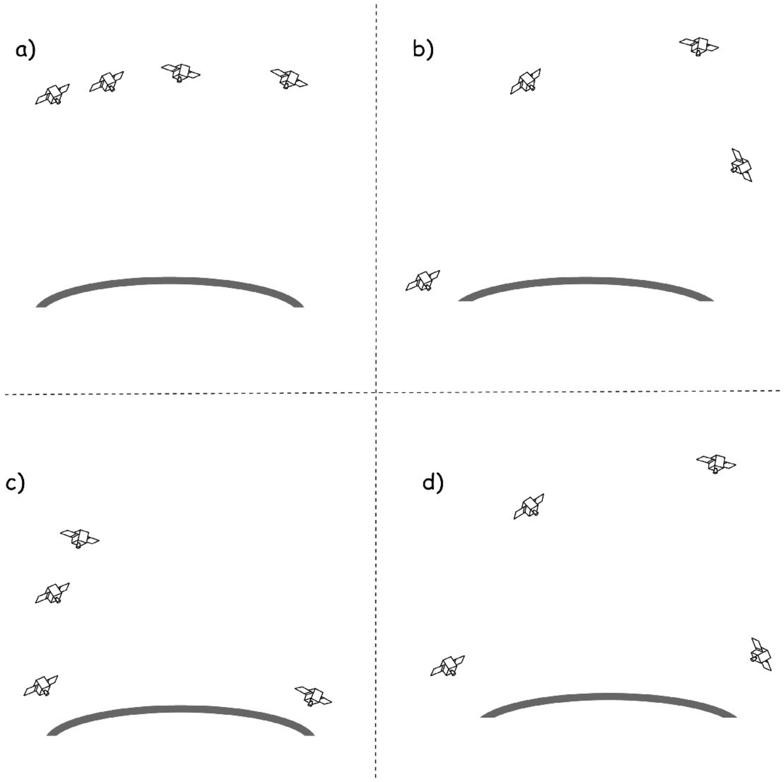
5.4 - How many satellites must you measure to obtain a three-dimensional position fix?

5.5 - What are the main sources and relative magnitudes of uncertainty in GNSS positioning?

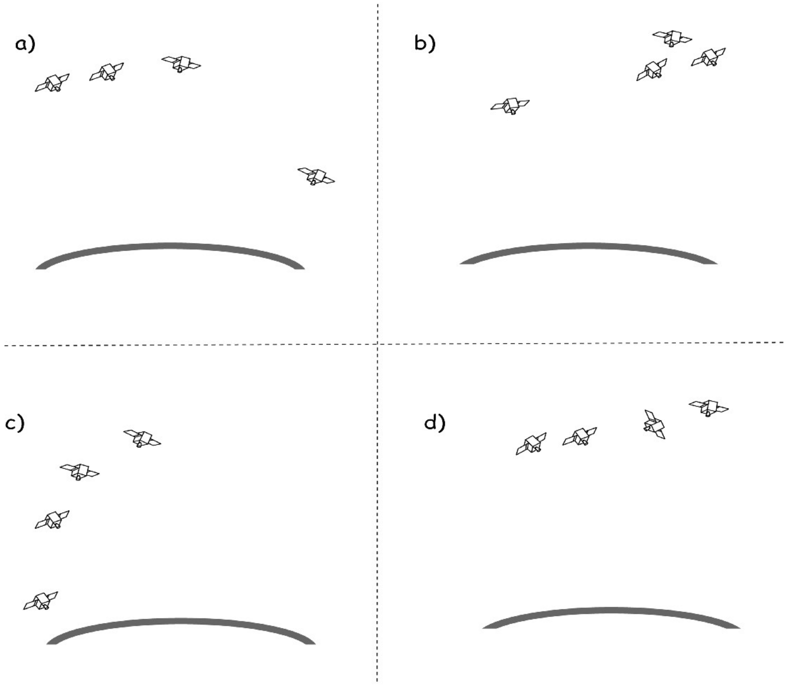
5.6 - How accurate is GNSS positioning? Be sure you specify a range, and describe under what conditions accuracies are at the high and low end of the range.

5.7 - What is a dilution of precision (DOP)? How does it affect GNSS position measurements?

5.8 - Which of the following figures depicts the lowest and highest PDOPs, assuming the observer is near the center of the drawn surface?



5.9 - What is the rank order, highest to lowest, for accuracy of GNSS data given the satellite constellations in the following figures, assuming the observer is near the center of the drawn surface?



5.10 - Describe the basic principle behind differential positioning.

5.11 - What are the differences between post processed and real-time differential positioning?

5.12 - What is the primary source of error reduced with a dual frequency GNSS receiver?

5.13 - Place these in order of accuracy, assuming the best equipment and practices applied to data collection:

a) dual frequency, real-time kinematic positioning,

b) precise point processing, and

c) post-processed, dual frequency positioning.

5.14 - How is GNSS accuracy affected by the local terrain horizon? How is it affected by canopy cover or building obstructions? Why does positional accuracy change as these conditions change?

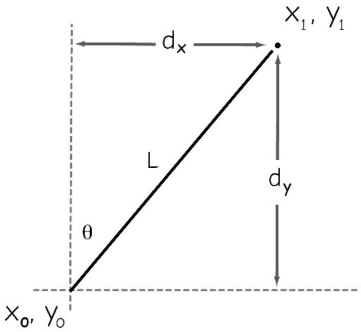
5.15 - How are GNSS data accuracy and efficiency (points collected per given time interval) related when collecting data in obstructed environments? Why? How is this controlled by field personnel?

5.16 - What is WAAS? Is it better or worse than ground-based differential positioning?

5.17 - Why are distance measurements devices and offset used when collecting GNSS data?

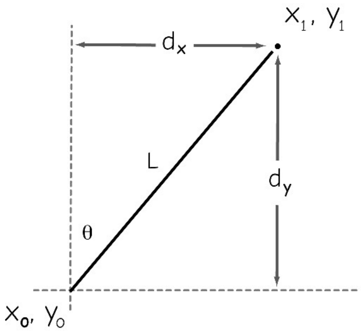
5.18 - What is COGO?

5.19 - Complete the table below, calculating missing elements according to the formulas presented in this chapter. Distances are in meters, the angle θ in degrees.



x_0	y_0	θ	L	d_x	d_y	x_1	y_1
10	20	30	500	250.0	433.0	260.0	453.0
400	97	60	1012	876.4	506.0		
937	12	84	1524	1515.7			
1540	1088	45	85				496.7
369	280	10	220				

5.20 - Complete the table below, calculating missing elements according to the formulas presented in this chapter. Distances are in meters, the angle θ in degrees.



x_0	y_0	θ	L	d_x	d_y	x_1	y_1
0	0	70	100	94.0	34.2	94.0	34.2
15	35	15	130	33.6	125.6		
400	0	45	200	141.4			
150	80	66	20				88.1
10	25	88	12				

5.21 - Fill in the missing cells, converting points between azimuths and bearings.

Point	1	2	3	4	5	6	7
Azimuth	138°			301°			18°14'22"
Bearing		S12°W	N33°E		S49°E	N88°W	

5.22 - Fill in the missing cells, converting points between azimuths and bearings.

Point	1	2	3	4	5	6	7
Azimuth	278°	42°		199°		245°	108°14′22″
Bearing	N82°W		S77°E		N1°W		

5.23 - Complete the table below for a traverse with the listed distances and bearings, given as azimuth degrees (drawing a rough sketch may help with the calculations). Note that many spreadsheet, calculator, and trigonometric functions require input in radians (approximately 57.2958 degrees = 1 radian).

Starting point P0, X = 10,128.3, Y = 6,096.4

Point ID	Azimuth	Distance	Delta X	Delta Y	X	Y
P1	32.4	122	65.4	103.0	10,193.7	6,199.4
P2	91.7	207	206.9	-6.1	10,400.6	6,193.3
P3	123.3	305				
P4	212.5	193				
P5	273.9	206				
P6	314.0	302			10,129.0	6,086.8

5.24 - Complete the table below for a traverse with the listed distances and bearings, given as azimuth degrees (drawing a rough sketch may help with the calculations). Note that many spreadsheet, calculator, and trigonometric functions require input in radians (approximately 57.2958 degrees = 1 radian).

Starting point P0, X = 1,200 Y = 400

Point ID	Azimuth	Distance	Delta X	Delta Y	X	Y
P1	95	105	104.6	-9.2	1,304.6	390.8
P2	192	77	-16.0	-75.3		
P3	262	204				
P4	6	104				
P5	18	33				
P6	105	88			1,192.6	399.2

5.25 - Complete the table for points 2 through 5, which have the listed GNSS-determined latitudes, longitudes, and ellipsoidal heights. The NGS data sheets for control locations near the measured points are listed, and you may use these to look up appropriate geoidal data.

Point (and nearest NGS point)	1 PID SF1124, French, Maine	2 PID ED1439 Rabun 2, GA	3 PID DF6074 Ellington N, WI	4 PID OW0320 Dennison, WY	5 PID RD1910 River RM3, OR
Latitude	47° 17′ 10.4″	34° 57′ 58.2″	44° 24′ 13.9″	43° 36′ 45.1″	45° 28′ 29.5″
Longitude	68° 19′ 1.9″	83° 17′ 55.2″	88° 34′ 40.9″	109° 29′ 43.9″	123° 50′ 31.0″
Ellipsoidal Height (m)	214.5	1244.1	188.4	2,594.5	-19.9
Geoidal Height (m)	-23.92	-29.15			
Orthometric Height (m)	238.42				

5.26 - Complete the table for points 2 through 5, which have the listed GNSS-determined latitudes, longitudes, and ellipsoidal heights. The NGS data sheets for control locations near the measured points are listed, and you may use these to look up appropriate geoidal data.

Point (and nearest NGS point)	1 PID DO4877, Cardinal, Minn.	2 PID AB6460 AUS APB3, TX	3 PID DY2143 Sta. Catal. CA	4 PID PX0445 Venus, WY	5 PID DL6000 Chel, Wash.
Latitude	45° 05′ 45.4″	30° 17′ 29.7″	33° 24′ 15.9″	44° 00′ 33.8″	47° 49′ 55.1″
Longitude	93° 00′ 17.9″	97° 41′ 34.3″	118° 24′ 54.2″	109° 30′ 20.2″	119° 59′ 21.6″
Ellipsoidal Height (m)	252.7	141.5	405.9	3604.5	382.1
Geoidal Height (m)	-27.45	-25.95			
Orthometric Height (m)	280.15				

6 Aerial and Satellite Images

Introduction

Aerial and satellite images are a valuable and common source of data for GIS. These images are data recorded from a distance; thus, photos and satellite images are often referred to as *remotely sensed* data. Remotely sensed data come in many forms; however, in the context of GIS we usually use the term to describe *aerial images* taken from aircraft using film or digital cameras, or *satellite images* recorded with satellite scanners. Until the 1970s, most mapping images were taken with film and aerial cameras. Digital aerial cameras are now a primary source of images and have replaced most film cameras. In addition, satellite scanners covering a range of resolutions are finding wide use. Whatever their origin, images are a rich source of spatial information and have been used as a basis for mapping for more than seven decades.

Remotely sensed images are valuable sources of spatial data for many reasons, including:

Broad-area coverage – images capture data from large areas at a relatively low cost and in a uniform manner (Figure 6-1). For example, it would take months to collect enough ground survey data to accurately produce a topographic map for 10 km^2. Images of a region this size may be collected in a few minutes and the topographic data extracted and interpreted in a few weeks.

Extended spectral range – photos and scanners can detect light from wavelengths outside the range of human eyesight. Some kinds of aerial photographs are sensitive to infrared wavelengths, a portion of the light spectrum that the human eye cannot sense. Aerial and satellite scanners sense even broader spectral ranges, up to thermal wavelengths and beyond. This expanded spectral range allows us to detect features or phenomena that appear invisible to the human eye.

Geometric accuracy – remotely sensed data may be converted to geometrically accurate spatial data. Aerial images are the source of many of our most accurate large-area maps. Under most conditions, aerial images contain geometric distortion due to imperfections in the camera, lens, or film systems, or due to camera tilt or terrain variation in the target area. Satellite scanners may also contain errors due to the imaging equipment or satellite platform. However, distortion removal methods are well established, and provide highly accurate spatial data from images. Cameras and imaging scanners have been developed specifically for the purpose of quantitative mapping. These systems are combined with techniques for identifying and removing most of the spatial error in aerial or satellite images, so spatially accurate data may be collected from images.

Permanent record – an image is fixed in time, so the conditions at the time of the photograph may be analyzed many years hence. Comparison of conditions over multiple dates, or determination of conditions at a specific date in the past are often quite valuable, and remotely sensed images are often the most accurate source of historical information.

Figure 6-1: Images are a valuable source of spatial data. The upper image, centered on northeastern Egypt, illustrates the broad-area coverage provided by satellite data. The lower image of pyramids in Egypt illustrates the high spatial detail that may be obtained (courtesy NASA, top, and Space Imaging, bottom).

Basic Principles

The most common forms of remote sensing are based on reflected electromagnetic energy. When energy from the sun or another source strikes an object, a portion of the energy is reflected. Different materials reflect different amounts of incoming energy, and this differential reflectance gives objects a distinct appearance. We use these differences to distinguish among objects.

Light is the principal energy form detected in remote sensing for GIS. Light energy is characterized by its *wavelength*, the distance between peaks in the electromagnetic stream. Each "color" of light has a distinctive wavelength, for example, we perceive light with wavelengths between 0.4 and 0.5 micrometers (μm) as blue. Light emitted by the sun is composed of several different wavelengths, and the full range of wavelengths is called the *electromagnetic spectrum.* A graph of this spectrum may be used to represent the amount of incident light energy across a range of wavelengths (Figure 6-2). The amount of energy in each wavelength is typically plotted against each wavelength value, yielding a curve depicting total electromagnetic energy reaching any object. Notice in Figure 6-2 that the amount of energy emitted by the sun increases rapidly to a maximum between 0.4 and 0.7 μm, and drops off at higher wavelengths. Some regions of the electromagnetic spectrum are named: X-rays have wavelengths of approximately 0.0001 μm, visible light is between 0.4 and 0.7 μm, and near-infrared light is between 0.7 and 1.1 μm.

Our eyes perceive light in the visible portion of the spectrum, between 0.4 and 0.7 μm. We typically identify three base colors: blue, from approximately 0.4 to 0.5 μm, green from 0.5 to 0.6 μm, and red from 0.6 to 0.7 μm. Other colors are often described as a mixture of these three colors at varying levels of brightness. For example, an equal mixture of blue, green, and red light at a high intensity is perceived as "white" light. This same mixture but at lower intensities produces various shades of gray. Other colors

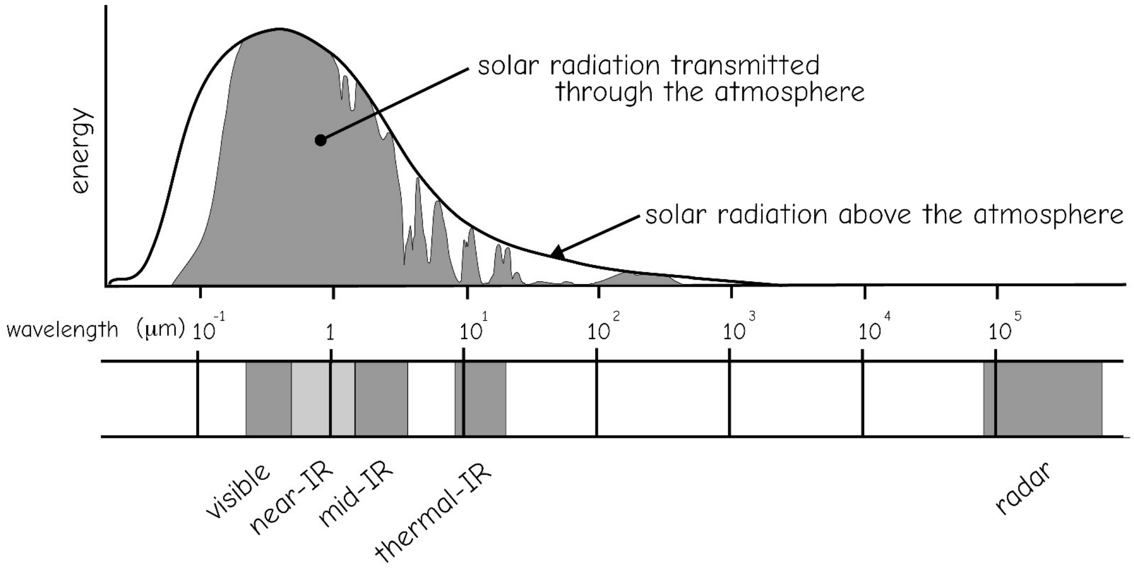


Figure 6-2: Electromagnetic energy is emitted by the sun and transmitted through the atmosphere (upper graph). Solar radiation is partially absorbed as it passes through the atmosphere. This results in variable surface radiation in the visible and infrared (IR) wavelength regions (lower graph).

are produced with other mixes; for example, equal parts red and green light are perceived as yellow. The specific combination of wavelengths and their relative intensities produce all the colors visible to the human eye.

Electromagnetic energy striking an object is reflected, absorbed, or transmitted. Most solid objects absorb or reflect incident electromagnetic energy and transmit none. Liquid water and atmospheric gasses are the most common natural materials that transmit light energy as well as absorb and reflect it.

Energy transmittance through the atmosphere is most closely tied to the amount of water vapor in the air. Water vapor absorbs energy in several portions of the spectrum, and higher atmospheric water content results in lower transmittance. Carbon dioxide, other gases, and particulates such as dust also contribute to atmospheric absorption, attenuating radiation in portions of the electromagnetic spectrum (Figure 6-2).

Most remote sensing systems are *passive*, in that they use energy generated by the sun and reflected off of the target objects. Aerial images and most satellite data are collected using passive systems. The images from these passive systems may be affected by atmospheric conditions in multiple ways. Figure 6-3 illustrates the many paths by which energy reaches a remote sensing device. Note that only one of the energy paths is useful, in that only the surface reflected energy provides information on the features of interest. The other paths result in no or only diffuse radiation reaching the sensor, and provide little information about the target objects. Most passive systems are not useful during cloudy or extremely hazy periods because nearly all the energy is scattered and no directly reflected energy may reach the sensor. Most passive systems rely on the sun's energy, so they have limited use at night.

Active systems are an alternative for gathering remotely sensed data under cloudy or nighttime conditions. Active systems generate an energy signal and detect the energy returned. Differences in the quantity and direction of the returned energy are used to identify the type and properties of features in an image. Radar (radio detection and rang-

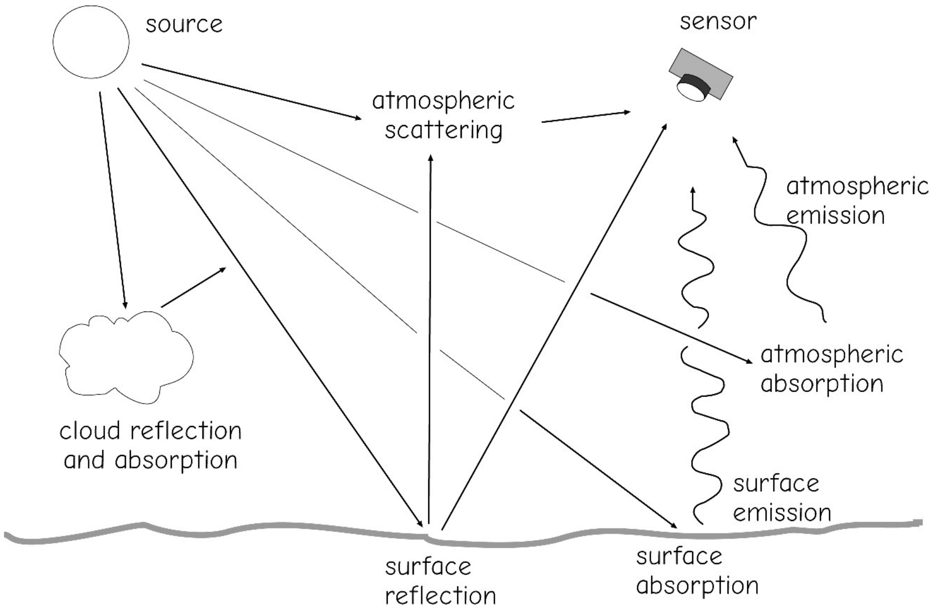


Figure 6-3: Energy pathways from source to sensor. Light and other electromagnetic energy may be absorbed, transmitted, or reflected by the atmosphere. Light reflected from the surface and transmitted to the sensor is used to create an image. The image may be degraded by atmospheric scattering due to water vapor, dust, smoke, and other constituents. Incoming or reflected energy may be scattered.

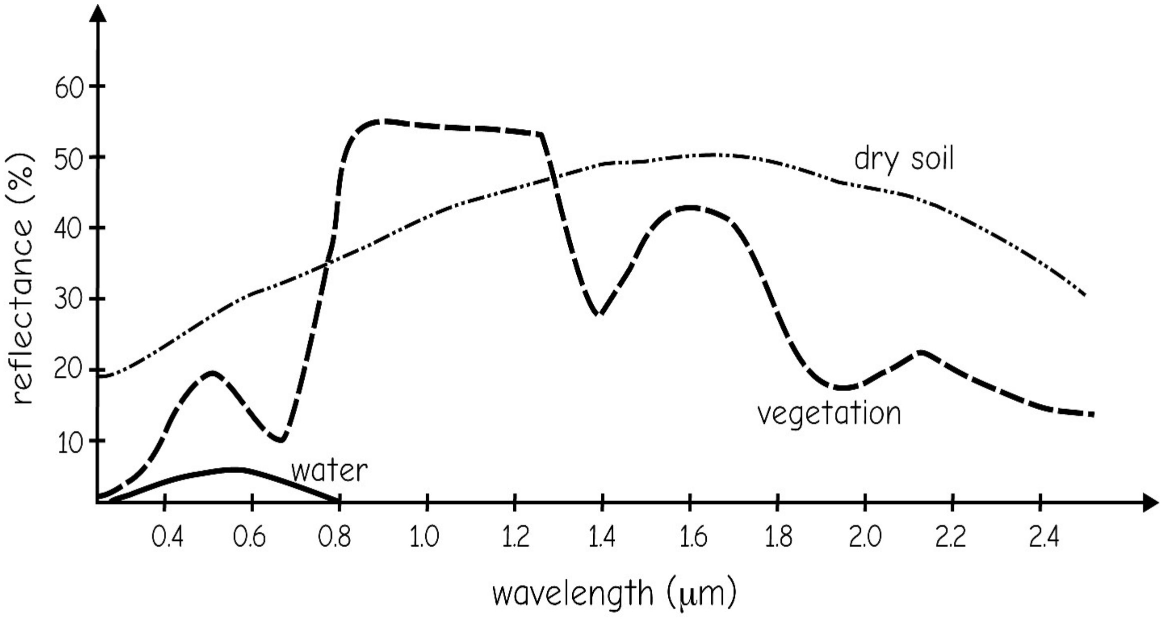


Figure 6-4: Spectral reflectance curves for some common substances. The proportion of incoming radiation that is reflected varies across wavelengths (adapted from Lillesand and Kiefer, 1999).

ing) is the most common active remote sensing system, while the use of LiDAR systems (light detection and ranging) is increasing. Radar focuses a beam of energy through an antenna, and then records the reflected energy. These signals are swept across the landscape, and the returns are assembled to produce a radar image. Because a given radar system is typically restricted to one wavelength, radar images are usually monochromatic (in shades of gray). These images may be collected day or night, and most radar systems penetrate clouds because water vapor does not absorb the relatively long radar wavelengths.

Natural objects appear to be the color they most reflect; for example, green leaves absorb more red and blue light and reflect more green light. Our eyes sense these differences in reflectance properties across a range of wavelengths to distinguish among objects. While we perceive differences in the visible wavelengths, these differences also extend into other portions of the electromagnetic spectrum that we cannot perceive (Figure 6-4). For example, individual leaves of many plant species appear to be the same shade of green; however, some reflect much more energy in the infrared portion of the spectrum, and thus appear to have a different "color" when viewed at infrared wavelengths.

Images we view either display a single range of wavelengths and appear in gray shades from black to white, or are "stacked" composite color images, with specific wavelength regions set to color blue, green, and red outputs (Figure 6-5). Color images are most common, as we usually collect at least three spectral bands, often four, and with some scanners up to hundreds of bands. Each band records the amount of energy in a specific set of wavelength, e.g., a near-infrared band may record the energy returned between 0.9 and 1.0 μm.

Each individual band may be thought of as averaging the energy returned over range of wavelengths. Bands may vary in limits, e.g., one camera may record a green band that averages from 0.6 to 0.68 μm, while another camera may average over the range from 0.62 to 0.68 μm. Data for a band may be conceived as stored in a grid, representing the reflectance for each cell, or pixel, over the area imaged.

We must assign each recorded band to a color on our displayed image. Images, both on a monitor or a hardcopy, are created by mixing primary colors, typically blue-green-red. We assign one layer recorded for one band to the green color, and variation in intensity sensed in the band translates into the range of dark to light greens assigned to the green layer. We do the same for red and blue layers, assigning an input band to each. We then combine the bands, that is, the colors for each corresponding pixel in the layers, to create a composite, multi-band image.

True color images are most common, and for these we assign the red-sensed band to the red color on the ouput, the green-sensed band to the green output color, and the blue-sensed band to the blue output color. This provides an image similar to the colors observed by the human eye for each feature.

We may also assign different band combinations or orders. Many aerial or satellite scanners sample in four bands, including an infrared band in addition to the three visible bands. The Sentinel satellite scanner, described later in this chapter, collects thirteen bands. We must choose which of these bands will be assigned respectively to the blue, green, and red display colors. Different band selections usually result in different displayed images, with abnormal colors for features, e.g., green vegetation may appear purple or red, depending on the band combinations used. We typically choose the bands that best help us identify important features.

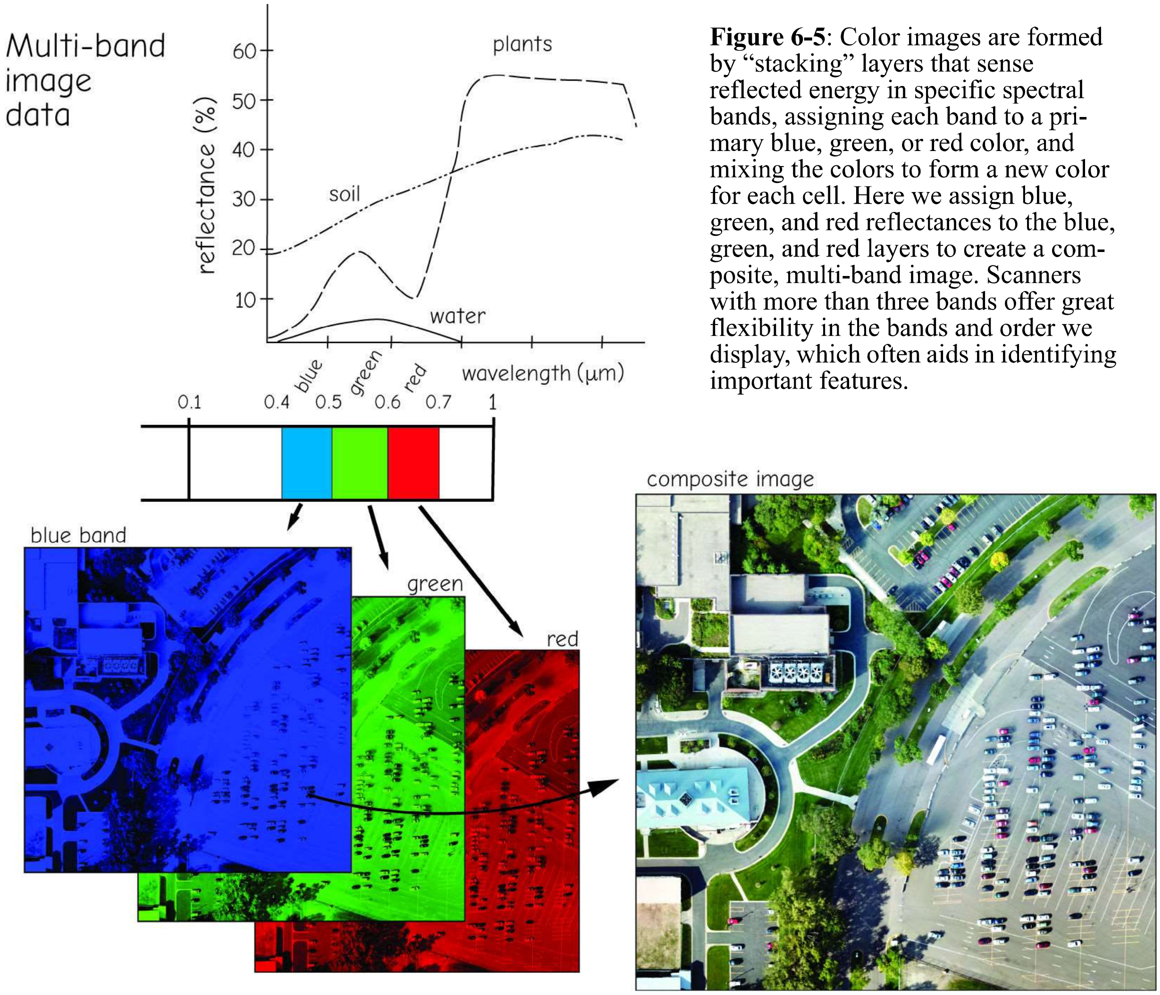


Figure 6-5: Color images are formed by "stacking" layers that sense reflected energy in specific spectral bands, assigning each band to a primary blue, green, or red color, and mixing the colors to form a new color for each cell. Here we assign blue, green, and red reflectances to the blue, green, and red layers to create a composite, multi-band image. Scanners with more than three bands offer great flexibility in the bands and order we display, which often aids in identifying important features.

Aerial Images

Images taken from airborne cameras are and have historically been a primary source of geographic data. Aerial photography quickly followed the invention of portable cameras in the mid-19th century, and became a practical reality with the development of dependable airplanes in the early 20th century (Figure 6-6). *Photogrammetry*, the science of measuring geometry from images, was well developed by the early 1930s, and there have been continuous refinements since. Aerial images underpin most large-area maps and surveys in most countries. Digital mapping cameras became common in the 21st century, largely supplanting aerial cameras, and are often carried aboard aircraft optimized for aerial surveys (Figure 6-7). Aerial images are routinely used in urban planning and management, construction, engineering, agriculture, forestry, wildlife management, and other mapping applications.

Figure 6-7: Aerial photographs are often taken from specialized aircraft, such as this low altitude airplane, or from helicopters or higher-flying, larger aircraft (courtesy Seabird Ltd.).

Although there are hundreds of applications for aerial images, most applications in support of GIS may be placed into three main categories. First, aerial images are often used as a basis for mapping, to measure and identify the horizontal and vertical locations of objects. Measurements on images offer a rapid and accurate way to obtain geographic coordinates, particularly when image measurements are combined with ground surveys. In a second major application, image interpretation may be used to categorize or assign attributes to surface features. For example, images are often used as the basis for landcover and infrastructure mapping, and to assess the extent of fire, flood, or other damage. Finally, images are often used as a backdrop for maps of other features, as when photographs are used as a background layer for soil survey maps produced by the U.S. National Resource Conservation Service.

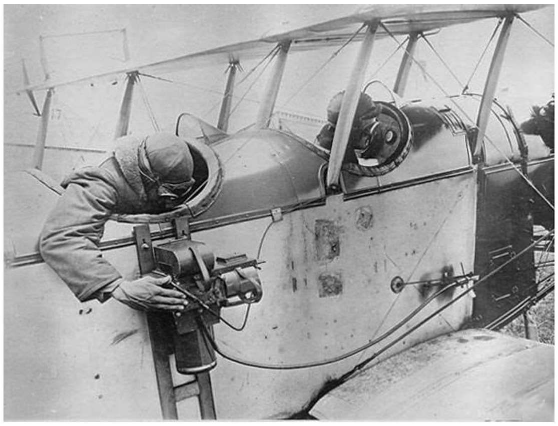


Figure 6-6: Aerial surveys began shortly after the development of reliable airplanes and portable, film-based cameras (courtesy Canadian Government Photographic Archives).

Camera Aircraft, Formats and Systems

Aerial camera systems are available that are specifically designed for mapping, so the camera and components are built to minimize geometric distortion and maximize image quality. Mapping cameras have features to reduce image blur due to aircraft motion, enhancing image quality. They maintain or record orientation angles, so distortions can be minimized. These camera systems are precisely made, sophisticated, highly specialized, and expensive, and most often used when large-area, high-resolution, accurate images are required.

Modern aerial cameras are typically mounted inside an aircraft, point through an underside bay (Figure 6-8). The camera mount and aircraft control systems are designed to maintain the camera optical axis as near vertical as possible. Aircraft navigation and control systems are specialized to support aerial photography, with precise positioning and flight control.

Aerial cameras specialized for spatial data collection are large, expensive, sophisticated devices, but in principle they are similar to simple cameras. A simple camera consists of a lens and a body (Figure 6-9). The lens is typically made of several individual glass elements, with a *diaphragm* or other mechanism to control the amount of light reaching the *sensing media*, the digital sensor or film that records light. These sensors have a characteristic dimension, sd, and for digital sensors, a pixel size, that when combined with the flying height (H), and focal length (h), determine the ground resolution and imaged area. An exposure control, such as a *shutter* within the lens, controls the length of time the sensing element is exposed to light. Cameras also have an *optical axis*, defined by the lens and lens mount. The optical axis is the central direction of the incoming image, and it is precisely oriented to intersect the sensor in a perpendicular direction. Digital sensors are connected to electronic storage so that successive images may be saved. Images are recorded at a surface called the camera's *focal plane*, perpendicular to the optical axis. The time, altitude, and other conditions or information regarding the photographs or mapping project may be recorded by the camera, often as an electronic *header* on digital image files, or on the *data strip* for film cameras, a line of text in the margin of the photograph.

Image scale and *extent* are important attributes of remotely sensed data. Image scale, as in map scale, is defined as the relative distance on the image to the correspond-

Figure 6-8: An example of the sophisticated system (upper left) for controlling digital image collection, here with a Leica Geosystems ADS40 digital aerial camera (lower right). These systems record and display flight paths and camera stations in real time, and may be used to plan, execute, and monitor image data collection (courtesy Leica Geosystems).

ing distance on the ground. For example, 1 inch on a 1:15,840-scale photograph corresponds to 15,840 inches on the Earth's surface. As shown in Figure 6-9, image scale will be h/H, the ratio of focal length to flying height.

Image extent is the area covered by an image, and depends on the physical size of the sensing area or element (sd in Figure 6-9), the camera focal length (h), and the flying height (H), according to:

$$gd = sd * H / h \qquad (6.1)$$

The extent depends on the physical size of the recording media, sd (e.g., 5 x 5 cm digital sensor), and the lens system and flying height. For example, a 5 cm sensing element with a 4 cm focal length lens flown at 3000 m height (about 10,000 ft) results in an extent of approximately 3.75 by 3.75 km^2 (5.1 mi^2) on the surface of the Earth.

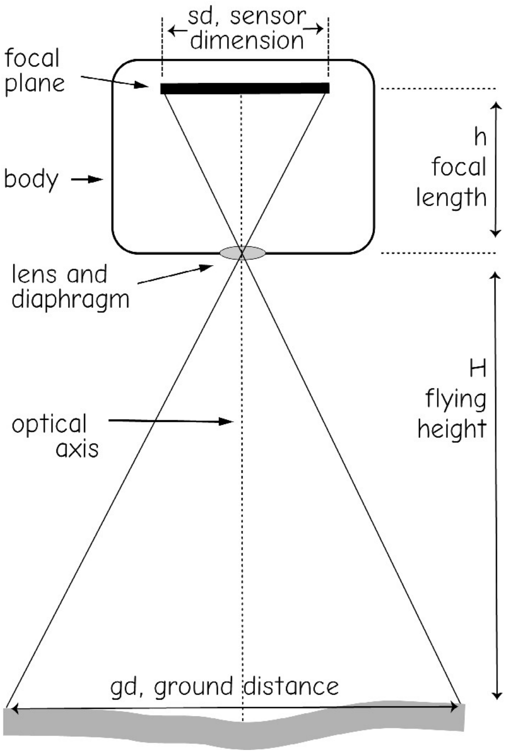


Figure 6-9: A simple camera.

Image resolution is another important concept. The resolution is the smallest object that can reliably be detected on the image. Resolution in digital cameras is often set by the pixel size, the size of individual sensing elements in an array. For example, a 5 x 5 cm array with 7,000 cells in each direction will have a cell size of 0.05/7,000, which is 7.1×10^{-6} m, or 7.13 µm.

The realized or ground resolution on aerial images may be approximately calculated from equation (6.1), substituting cell dimension for sensor dimension, sd. In our example, if the camera has a 10 cm (0.1 m) focal length, and is flown at 3,000 m, the ground resolution is:

$$0.21 \text{ m} = 7.1 \times 10\text{-}6 * 3{,}000 / 0.1 \qquad (6.2)$$

Resolution in aerial photographs is more complicated, and depends on film grain size and exposure properties, and is often tested via photographs of alternating patterns of black and white lines. At some threshold of line width, the difference between black and white lines cannot be distinguished, giving the effective resolution.

Digital Aerial Cameras

Digital aerial cameras are the most common systems used for aerial mapping, and routinely provide high-quality images. Film cameras were most common for the 1920s through the mid-1990s, but we have transitioned to digital cameras. Digital aerial cameras provide many advantages over film cameras, including greater flexibility, easier planning and execution, greater stability, and direct to digital output. While film cameras are still in use today, camera production has effectively ceased, and film will soon follow.

Digital cameras typically consist of an electronic housing that sits atop a lens assembly (Figure 6-10). The lens focuses light onto charge-coupled devices (CCDs) or similar electronic scanning elements. A

Figure 6-10: Digital aerial cameras are superficially similar to film aerial cameras, but typically contain many and more sophisticated electronic components (courtesy Leica Geosystems).

CCD is a rectangular array of *pixels*, or picture elements, that respond to light.

The CCD is composed of layers of semiconducting material with appropriate reflective and absorptive coatings, insulators, and conducting electrodes (Figure 6-11). Incoming radiation passes through the coatings and into the semiconductors, dislodging electrons and creating a voltage or current. Response may be calibrated and converted to measures of light intensity. Response varies across wavelength, but can be tuned to wavelength regions by manipulating semiconductor composition. Since the pixels are in an array, the array then defines an image.

Digital cameras sometimes use a multi-lens cluster rather than a single lens, or they may split the beam of incoming light via a prism, diffraction grating, or some other mechanism (Figure 6-12). Since CCDs are typically configured to be sensitive to only a narrow band of light, multiple CCDs may be used, each with a dedicated lens and a specific waveband. Multiple CCDs typically allow more light for each pixel and waveband, but this increases the complexity of the camera system. If a multi-lens system is

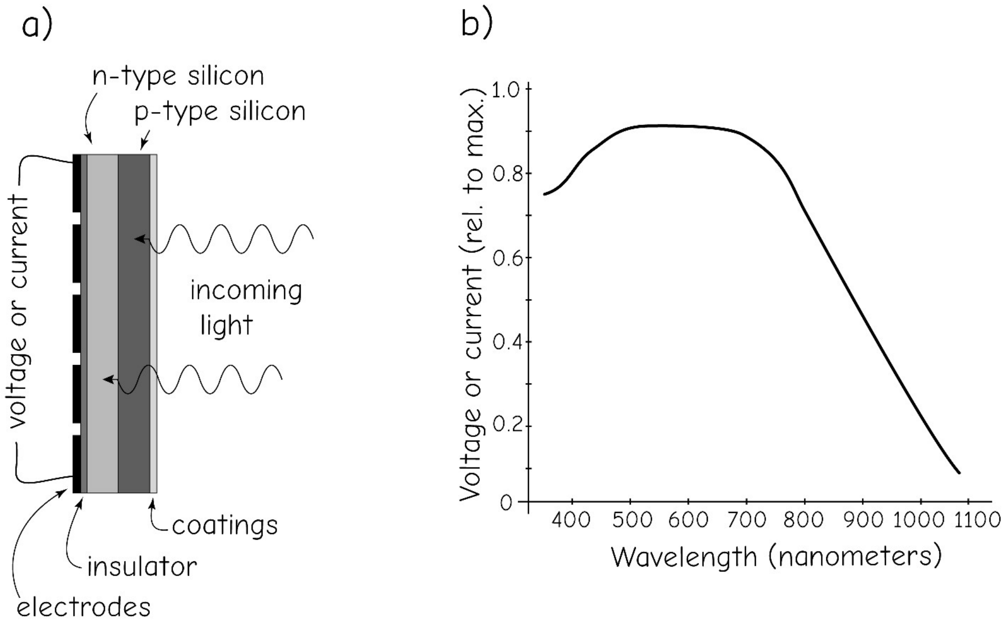


Figure 6-11: CCD response for a typical silicon-based receptor. The CCD is a sandwich of semiconducting layers (a, on left) that generates a current or voltage in proportion to the light received. Response varies over a wavelength region (b, on right).

used, the individual bands from the multiple lenses and CCDs must be carefully *coregistered*, or aligned, to form a complete multiband image.

Digital cameras most commonly collect images in the blue (0.4–0.5 µm), green (0.5–0.6 µm), or red (0.6–0.7 µm) portion of the electromagnetic spectrum. This provides an image approximately equal to what the human eye perceives. Systems may also record near-infrared reflectance (0.7–1.1 µm), particularly for vegetation mapping. The camera may also have a set of filters that may be placed in front of the lens, for example, for protection or to reduce haze.

Digital cameras typically have a computer control system, used to specify the location, timing, and exposure; record GPS and aircraft altitude and orientation information; provide data transfer and storage; and allow the operator to monitor progress and image quality during data collection.

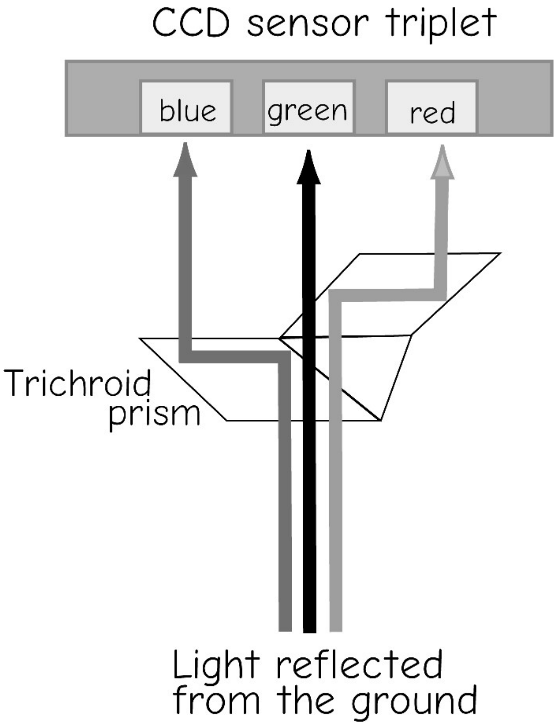


Figure 6-12: Digital cameras often use a prism or other mechanism to separate and direct light to appropriate CCD sensors (adapted from Leica Geosystems).

Digital cameras may have several features to improve data quality. For example, digital cameras may employ electronic image motion compensation, combining information collected across several rows of CCD pixels. This may lead to sharper images while reducing the likelihood of camera malfunction due to fewer moving parts. In addition, digital data may be recorded in long, continuous strips, easing the production of image mosaics.

Film and Film Cameras

While nearly all future aerial images will be collected with digital cameras, there is a vast archive of past aerial images collected on film.These images come in various *formats*, or sizes, usually specified by the edge dimension, for example, 240 mm (9 in) on a side. These large-format films were most common for mapping.

Film consists of a sandwich of light-sensitive coatings spread on a plastic sheet. Black and white films have a single layer of light-sensitive material while color films have several layers. Each layer is sensitive to a different set of wavelengths. These layers, referred to as the *emulsions*, undergo chemical reactions when exposed to light. More light energy falling on the film results in a more complete chemical reaction, and hence a greater film exposure.

Films may be categorized by the wavelengths of light they respond to. Black and white films are sensitive to light in the visible portion of the spectrum, from 0.4 to 0.7 µm, and are often referred to as *panchromatic* films. *True color* film is also sensitive to light across the visible spectrum, but in three separate colors.

Infrared films have been developed and were widely used when differences in vegetation type were of interest. These films are sensitive through the visible spectrum and longer infrared wavelengths, up to approximately 0.95 µm.

Lens and Camera Distortion

The camera and lens system may be a significant source of geometric error in aerial images. The perfect lens–camera–detector system would exactly project the viewing geometry of the target onto the image recording surface, either film or CCD. The relative locations of features on the image in a perfect camera system would be exactly the same as the relative locations on any viewing plane that is in front of the lens. Real camera systems are not perfect and may distort the image. For example, the light from a point may be bent slightly when traveling through the lens, or the film may shrink or swell, both causing a distorted image.

Radial lens displacement is a common form of distortion that should be corrected in most camera systems used to develop GIS data (Figure 6-13). All manufactured lenses contain imperfections in the lens surfaces. These cause a radial displacement, either inward or outward, from the true image location. Radial lens displacement is typically quite small in mapping camera systems, but it may be quite large in other systems. A radial displacement curve is often developed for a mapping camera lens, and this curve may be used to correct radial displacement errors when the highest mapping accuracy is required.

Mapping camera systems are engineered to minimize systematic errors. Lenses are designed and precisely manufactured so that image distortion is minimized. Lens mountings, the detectors, and the camera body are optimized to ensure a faithful rendition of image geometry. Films are designed so that there is limited distortion under tension on the camera spools. This optimization leads to extremely high geometric fidelity in the camera/lens system. Thus, camera and lens distortions in mapping cameras are typically much smaller than other errors, for example, tilt and terrain errors, or errors in converting the image data to forms useful in a GIS.

Camera-caused geometric errors may be quite high when a non-mapping camera is used, such as when photographs are taken with non-mapping digital cameras. Radial distortion may be extreme, and these systems are likely to have large geometric errors when compared to mapping cameras. Non-mapping cameras may be used, either when geometric accuracy requirements are low, or when proper methods are used to calibrate and correct the geometric errors. Software has been developed with the specific purpose of error removal through image calibration, distortion removal, and subsequent processing into planar mosaics. If these software and methods are not applied, the geometric quality of any non-mapping camera system should be evaluated prior to use in a mapping project.

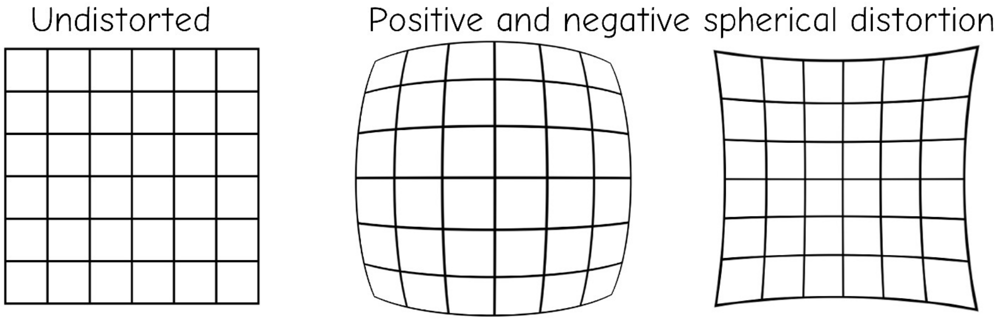


Figure 6-13: An example of radial lens distortion, common in most lenses, and quite large in lenses not designed for aerial image acquisition. The left-most figure shows a perfect lens, while the center and right figures show spherical or “barrel” distortion outward and inward, respectively. Lenses must be carefully crafted to minimize this distortion, and then calibrated so that remaining distortion may be removed through subsequent image processing.

Small Unmanned Aerial Vehicles: Drones

Small, unmanned planes and helicopters have been introduced over the past decade (Figure 6-14). Variously called unmanned aerial vehicles (UAVs), remotely piloted vehicles (RPVs), or simply drones, they may substantially reduce the cost and increase the flexibility of data collection. Data may also require increased processing times and exhibit more variable accuracy, given the small footprint and greater difficulty maintaining a level orientation in these small aircraft. They most often carry small digital cameras, although these cameras are often not optimized for minimizing geometric distortion. UAVs may also carry lidar and other sensor systems. Typically, image acquisitions are at low flying heights, close ranges, extremely high resolutions (inches to 10's of centimeters), and small areas, and distortion removed to variable levels through software.

One primary advantage of drones is low cost and ease of deployment. Many drone systems for GIS data collection currently cost less than $20,000, and some for below $5,000, including all subsystems for converting raw images into orthographic, georeferenced images, and point clouds, which are dense collections of 3D point measurements. Drones may be carried to a site and launched, often by hand for the smallest units, using preprogrammed flight lines to collect images along a specified path (Figure 6-15). This allows data collection at the time and place of interest, provided weather and other conditions are apt. Smaller data collection windows may be utilized, increasing the likelihood of data acquisition.

Figure 6-14: The senseFly eBee, an example of small UAV/drones optimized for spatial data collection (courtesy senseFly Ltd.).

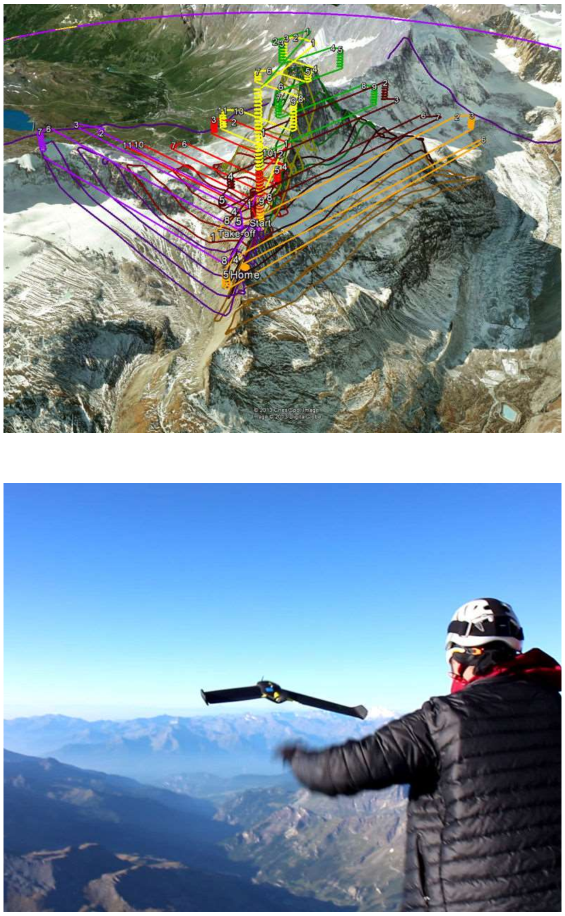


Figure 6-15: UAV systems for GIS typically include flight-planning utilities to specify adequate image overlap and approximate positioning (courtesy senseFly Ltd.).

High data resolutions may be another advantage of UAVs. Because they may be flown at low altitudes, pixel sizes of a few centimeters or less are possible. Individual bridge beams, rooftop fans, or paths may be resolved, leading to more detailed data, with very high point densities (Figure 6-16).

Smaller UAVs are limited in their data collection rates, and likely will not be suitable for areas much larger than a few to tens of square kilometers. They fly at relatively

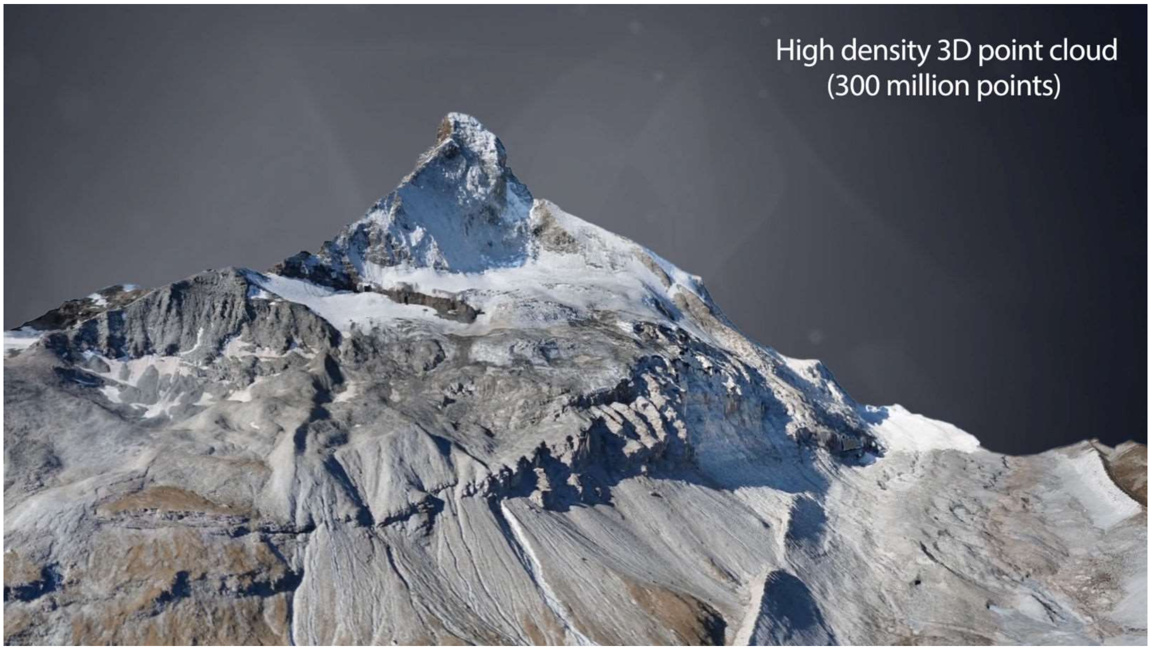


Figure 6-16: A reconstructed three-dimensional data set for the Matterhorn, in Switzerland, collected with a UAV. Point densities are selected, based on the time available and area to be sampled, and not fixed, as with many satellite and in some cases aerial systems (courtesy senseFly Ltd.).

low speeds, and typically carry small camera systems with commensurately small image footprints. UAVs for GIS data collection range in size from less than a meter wing-span through several to tens of meters, and there is a trade-off, with increasing costs associated with increasing system through-put. Larger UAVs may collect data at rates approaching current manned aerial systems, but in doing so lose many of the cost, flexibility, and resolution advantages.

Spatial data from small UAVs may be of more variable quality than that from larger aerial platforms, so care should be taken in choosing the proper system and evaluating realized spatial accuracies (Figure 6-17). Cameras and lens are typically not specifically designed for spatial accuracy, with greater spherical and other lens distortion, and small UAVs often deploy accurate GNSS and hence potentially larger positioning errors, and more dependent on less well-trained operators or analysts. Many of these potential limitations may be addressed in appropriately developed software. For example, image correlation and 3D reconstruction algorithms may be robust in finding correct image orientations, and advising the analysts when they are unable to reach acceptable accuracies. Lens or other system distortions may be removed through precise calibration using standardized test patterns. These measures are not available or applied for all UAV systems or softwares marketed as spatial data collection tools. It is up to the data consumer to verify the proclaimed accuracy of any system.

Figure 6-17: An example of a small UAV optimized for geometric accuracy. Camera, lens, GNSS, and software systems have been optimized to provide the most accurate data possible at a modest cost. Each data collection system should be assessed against desired accuracies (courtesy senseFly Ltd.).

Spatial Accuracy of Aerial Images

Aerial images are a rich source of spatial information, but most aerial images contain geometric distortions (Figure 6-18). We most often work with planar spatial data layers that are *orthographic*, with all objects projected onto a common plane (Figure 6-19). Objects above or below the plane are vertically projected down or up onto the horizontal plane. Thus, the top and bottom of a building should be projected onto the same location in the datum plane. In our ideal data set, the tops of all buildings would be visible, and all building sides not. Except for overhangs, bridges, or similar structures, the ground surface is visible everywhere.

Unfortunately, most aerial or satellite images provide a non-orthographic *perspective view* (Figure 6-18 and Figure 6-19, left). Perspective views give a geometrically distorted image of the Earth's surface. Distortion affects the relative positions of objects, and uncorrected data derived from aerial images may not directly overlay data in an accurate orthographic map.The amount of distortion in aerial images may be reduced by selecting the appropriate camera, lens, flying height, and type of aircraft. Distortion may also be controlled by collecting images under proper weather conditions during periods of low wind and by employing skilled pilots and operators. However, some aspects of the distortion may not be controlled, and no camera system is perfect, so there is some geometric distortion in every uncorrected aerial image. The real question becomes, "is the distortion and geometric error below acceptable limits, given the intended use of the spatial data?" This question is not unique to aerial images; it applies equally well to satellite images, spatial data derived from GPS and traditional ground surveys, or any other data.

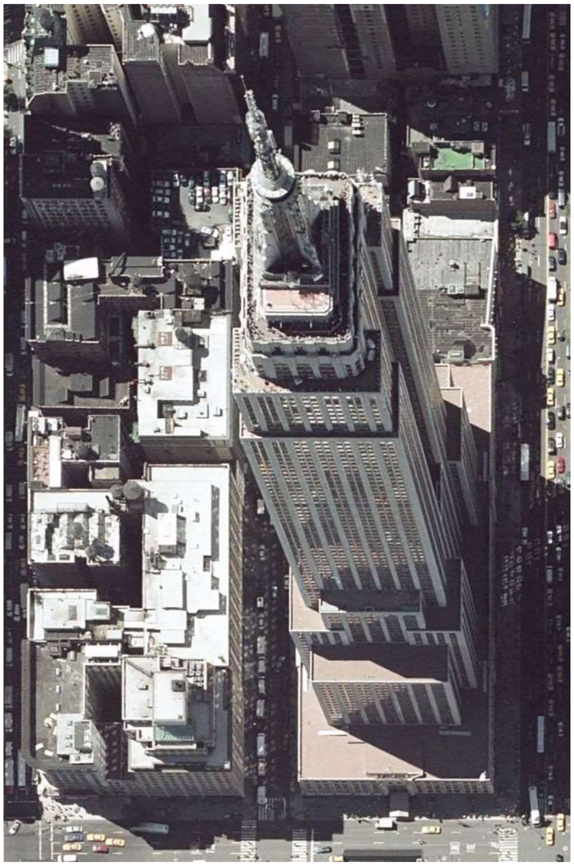


Figure 6-18: Tilt distortion is common on aerial and some satellite images, the result of perspective distortion when imaging the top and bottom of buildings, or any objects at different elevations.

The largest distortion in most aerial images comes from two sources: terrain variation and camera tilt, particularly when using an aerial mapping camera or when properly calibrating and correcting non-mapping cameras, e.g., with drone/UAV images. Atmospheric bending is relatively minor under most conditions when collecting aerial images, but may still be unacceptable, particularly when the highest-quality data are required. Established methods should be used to reduce the typically dominant tilt and terrain errors, and for non-mapping cameras, the previously described lens and camera distortion.

Terrain and Tilt Distortion in Aerial Images

Terrain variation, defined as differences in elevation within the image area, is often the largest source of geometric distortion in aerial images. Terrain variation causes *relief displacement*, defined as the radial displacement of objects that are at different elevations.

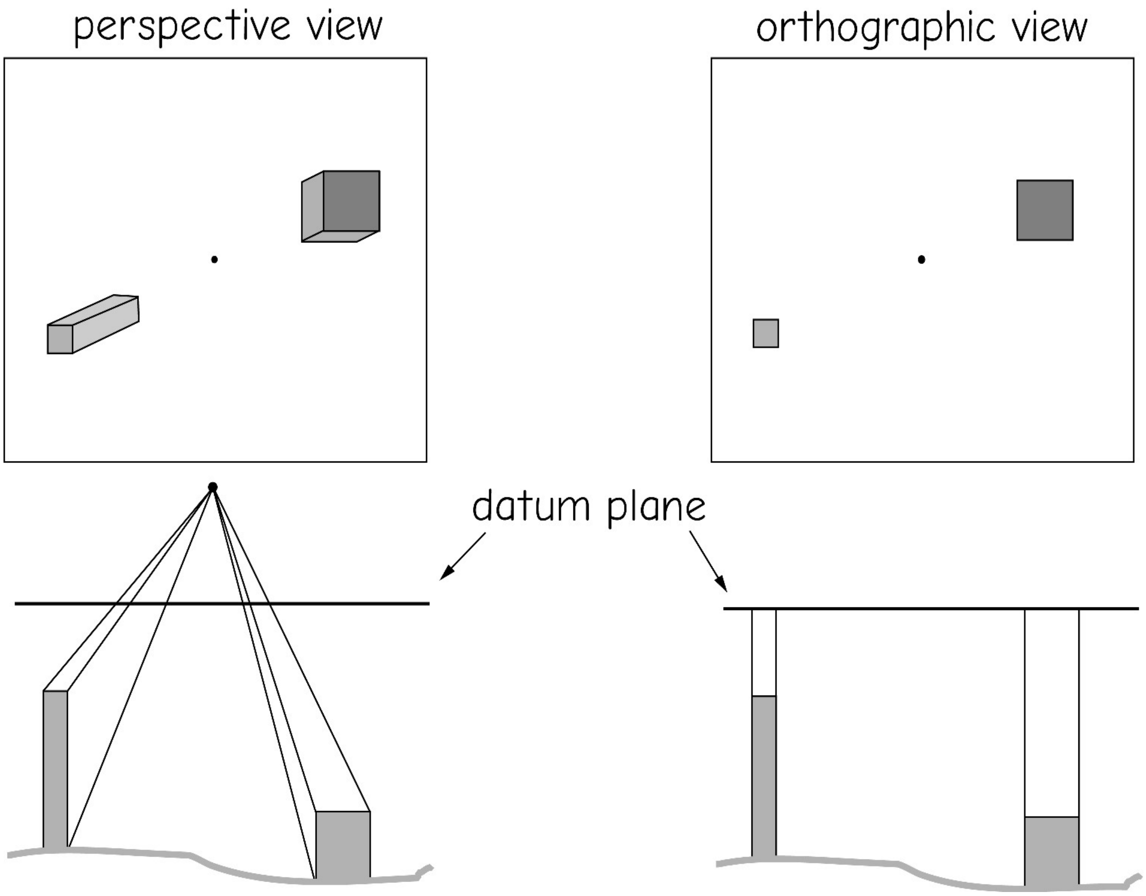


Figure 6-19: Orthographic (left) and perspective (right) views. Orthographic views project at right angles to the datum plane, as if viewing from an infinite height. Perspective views project from the surface onto a datum plane from a fixed viewing location.

Figure 6-20 illustrates the basic principles of relief displacement. The figure shows the photographic geometry over an area with substantial differences in terrain. The reference surface (datum plane) in this example is chosen to be at the elevation of the *nadir* point directly below the camera, N on the ground, imaged at n on the photograph. The camera station P is the location of the camera at the time of the photograph. We are assuming a vertical photograph, meaning the optical axis of the lens points vertically below the camera and intersects the reference surface at a right angle at the nadir location.

The locations for points A and B are shown on the ground surface. The corresponding locations for these points occur at A′ and B′ on the reference datum surface. These locations are projected onto the imaging sensor or film, as they would appear in a photograph taken over this varied terrain. In a real camera, the sensor is behind the lens; however, it is easier to visualize the displacement by showing the sensor in front of the lens, and the geometry is the same. Note that the points a and b are displaced from their reference surface locations, a′ and b′. The point a is displaced radially outward relative to a′, because the elevation at A is higher than the reference surface. The displacement of b is inward relative to b′, because B is lower than the reference datum.

Note that any points that have elevations exactly equal to the elevation of the reference datum will not be displaced, because the reference and ground surfaces coincide at those points.

Figure 6-20 illustrates the following key characteristics of terrain distortion in vertical aerial images:

Terrain distortions are radial – higher elevations are displaced outward, and lower elevations displaced inward relative to the center point.

Relief distortions affect angles and distances on an image – relief distortion changes the distances between points, and will change most angles. Straight lines on the ground will not appear to be straight on the image, and areas will expand or shrink.

Scale is not constant on aerial images – scale changes across the photograph and depends on the magnitude of the relief displacement. We may describe an average scale for a vertical aerial photograph over varied terrain, but the true scale between any two points will often differ.

A vertical aerial image taken over varied terrain is not orthographic – we cannot expect geographic data from terrain-distorted images to match orthographic data in a GIS. If the distortions are small relative to digitizing error or other sources of geometric error, then data may appear to match data from orthographic sources. If the relief displacement is large, it will add significant errors.

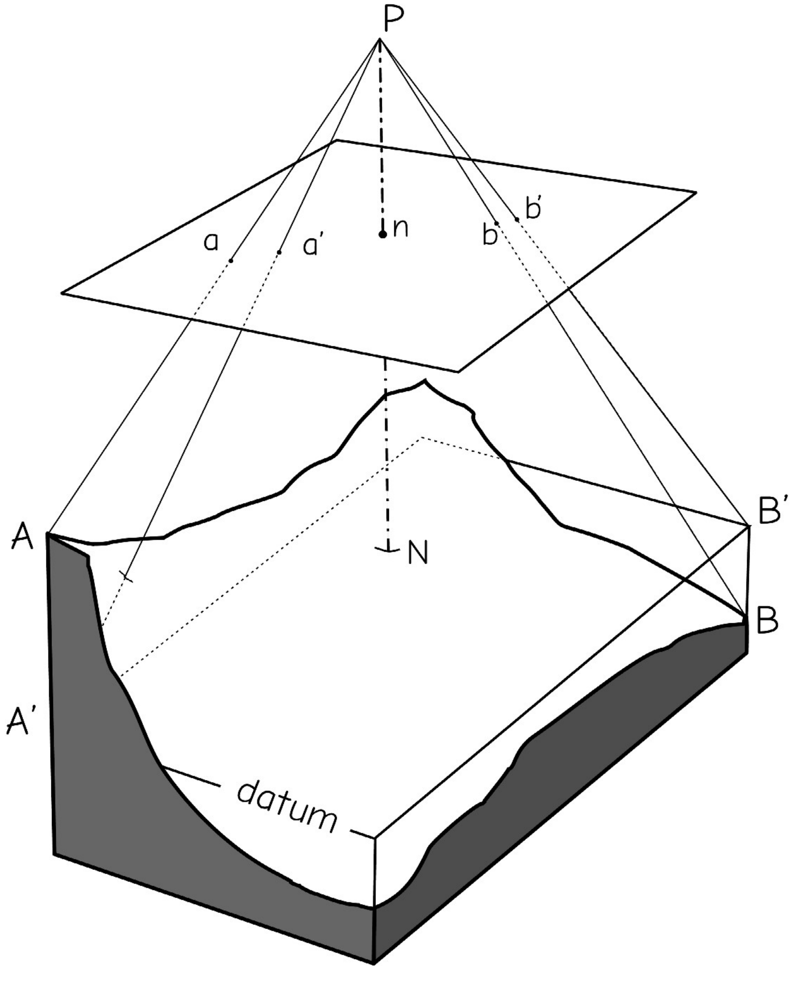


Figure 6-20: Geometric distortion on an aerial photograph due to relief displacement. P is the camera station, N is the nadir point. The locations of features are shifted, and the magnitude of their shift depends on their differences in elevation from the datum, and the point's distance from nadir. Unless corrected, this will result in non-orthographic images, and errors in location, distance, shape, and area for any spatial data derived from these images (adapted from Lillesand, Kiefer, and Chipman, 2007).

Camera tilt may be another large source of positional error in aerial images. Camera tilt, in which the optical axis points at a non-vertical angle, results in complex *perspective convergence* in aerial images (Figure 6-21). Objects farther away appear to be closer together than equivalently spaced objects that are nearer the observer (Figure 6-22). Tilt distortion is zero in vertical photographs, and increases as tilt increases.

Figure 6-21: An example of tilt convergence. Crop bands of equal width appear narrower towards the horizon in this highly tilted image (courtesy USDA).

Contracts for aerial mapping missions typically specify tilt angles of less than 3 degrees from vertical. Perspective distortion caused by tilt is somewhat difficult to remove, and removal tends to reduce resolution near the edges of the image. Therefore, efforts are made to minimize tilt distortion by maintaining a vertical optical axis when images are collected. Camera mounting systems are devised so the optical axis of the lens points directly below, and pilots attempt to keep the aircraft on a smooth and level flight path as much as possible. Planes have stabilizing mechanisms, and cameras may be equipped with compensating mechanics to

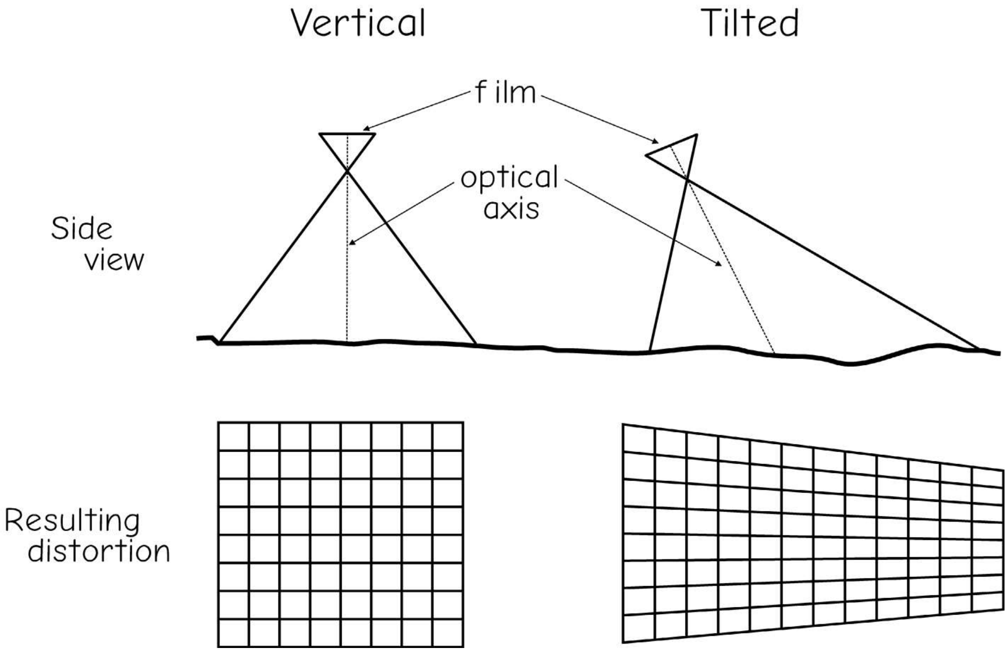


Figure 6-22: Image distortion caused by a tilt in the camera optical axis relative to the ground surface. The perspective distortion, shown at the bottom right, results from changes in the viewing distance across the photograph.

maintain an untilted axis. Despite these precautions, tilt happens, due to flights during windy conditions, pilot or instrument error, or system design.

Tilt is often characterized by three angles of rotation, often referred to as omega (ω), phi (ϕ), and kappa (κ). These are angles of rotation about the X, Y, and Z axes that define three-dimensional space (Figure 6-23). Rotation about the Z axis alone does not result in tilt distortion, because it occurs around the axis perpendicular with the surface. If ω and ϕ are zero, then there is no tilt distortion. However, tilt is almost always present, even in small values, so all three rotation angles are required to describe and correct it.

Tilt and terrain distortion may both occur on aerial images taken over varied terrain. Tilt distortion may occur even on vertical aerial images, because tilts up to 3 degrees are usually allowed. The overall level of distortion depends on the amount of tilt and the variation in terrain, and also on the photographic scale. Not surprisingly, errors increase as tilt or terrain increase, and as photographic scale becomes smaller.

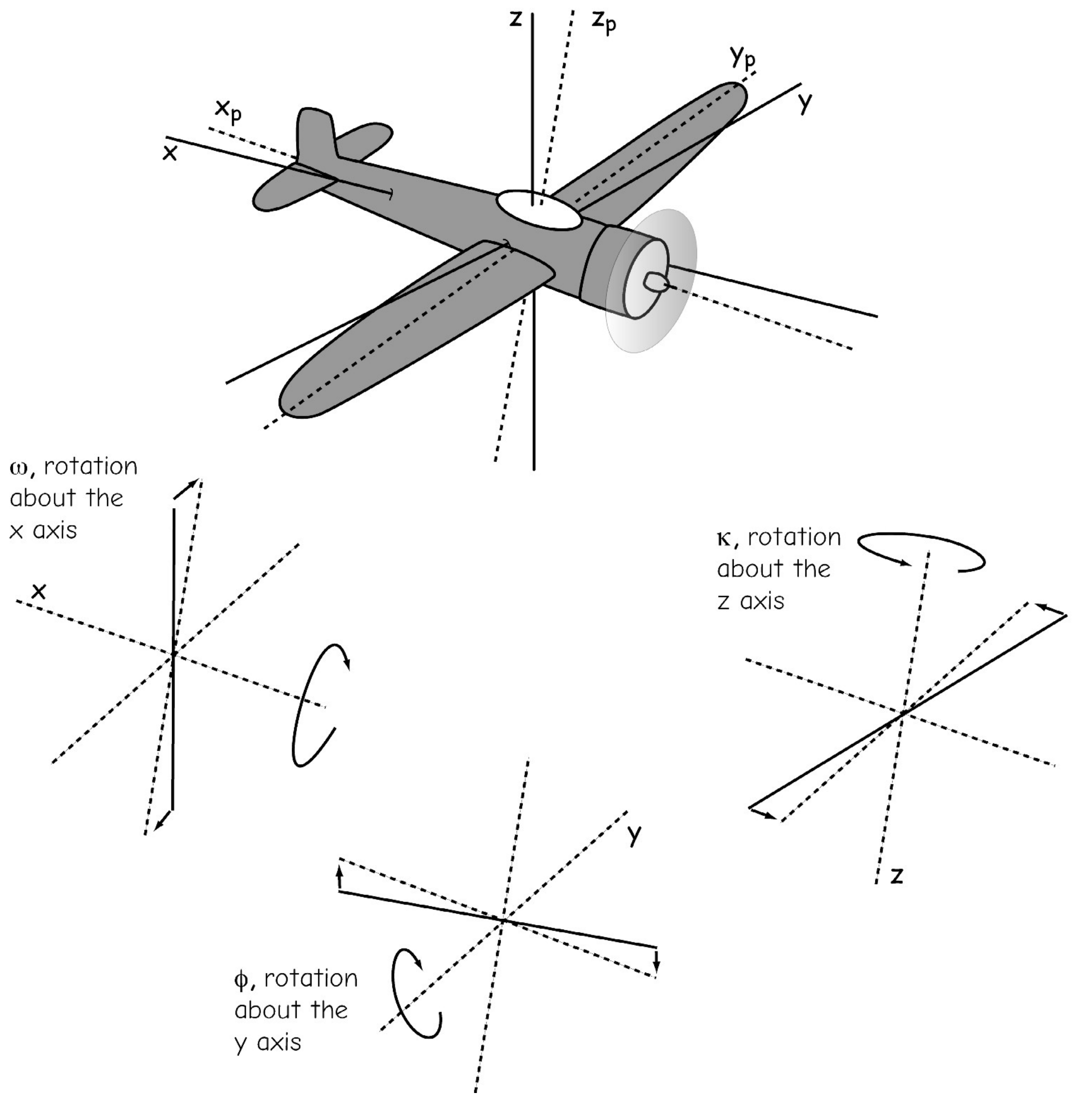


Figure 6-23: Image tilt angles are often specified by rotations about the X axis (angle ω), the Y axis (angle ϕ), and the Z axis (angle κ).

Figure 6-24 illustrates the changes in total distortion with changes in tilt, terrain, and image scale. This figure shows the error that would be expected in data digitized from vertical aerial images when only applying an affine transformation, a standard procedure used to register orthographic maps (see Chapter 4). The process used to produce these error plots mimics the process of directly digitizing from uncorrected aerial images. Note first that there is zero error across all scales when the ground is flat (terrain range is zero) and there is no tilt (bottom line, left panel in Figure 6-24). Errors increase as image scale decreases, shown by increasing errors as you move from left to right in both panels. Error also increases as tilt or terrain increase.

Geometric errors can be quite large, even for vertical images over moderate terrain (Figure 6-24, right side). These graphs clearly indicate that geometric errors will occur when digitizing from vertical aerial images, even if the digitizing system is perfect and introduces no error. Thus, the magnitude of tilt and terrain errors should be assessed relative to the geometric accuracy required before data capture. For most projects, errors are too large when digitizing from uncorrected aerial images, so some form of correction, often based on stereo coverage, is required.

Stereo Photographic Coverage

As noted above, relief displacement in vertical aerial images adds a radial displacement that depends on terrain heights. The larger the terrain differences, the larger the relief displacement. This relief displacement may be a problem if we wish to produce a map from a single photograph. Photogrammetric methods can be used to remove the distortion. However, if two overlapping photographs are taken, called a *stereopair*, then these photographs may be used together to determine the relative elevation differences. Relief displacement in a stereopair may be used to determine elevation and remove distortion. Many mapping projects collect *stereo photographic coverage*, in which sequential photographs in a flight line overlap, called *endlap*, and adjacent flightlines overlap, called *sidelap* (Figure 6-25). Stereo photographs typically have near 65% endlap and 25% sidelap. Some digital cameras collect data in continuous strips and so only collect sidelap. Drone-based collections often have nearly 100% endlap or sidelap, depending on the project and method for data collection.

A *stereomodel* is a three-dimensional perception of terrain or other objects that we see when viewing a stereopair. As each eye looks at a different, adjacent photograph

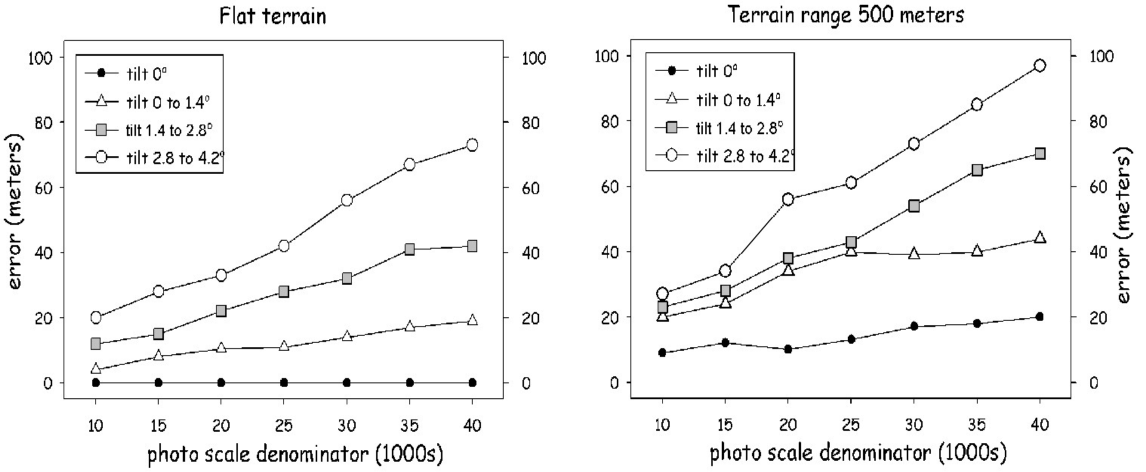


Figure 6-24: Terrain and tilt effects on mean positional error when digitizing from uncorrected aerial images. Distortion increases when tilt and terrain increase, and as photo scale decreases (from Bolstad, 1992).

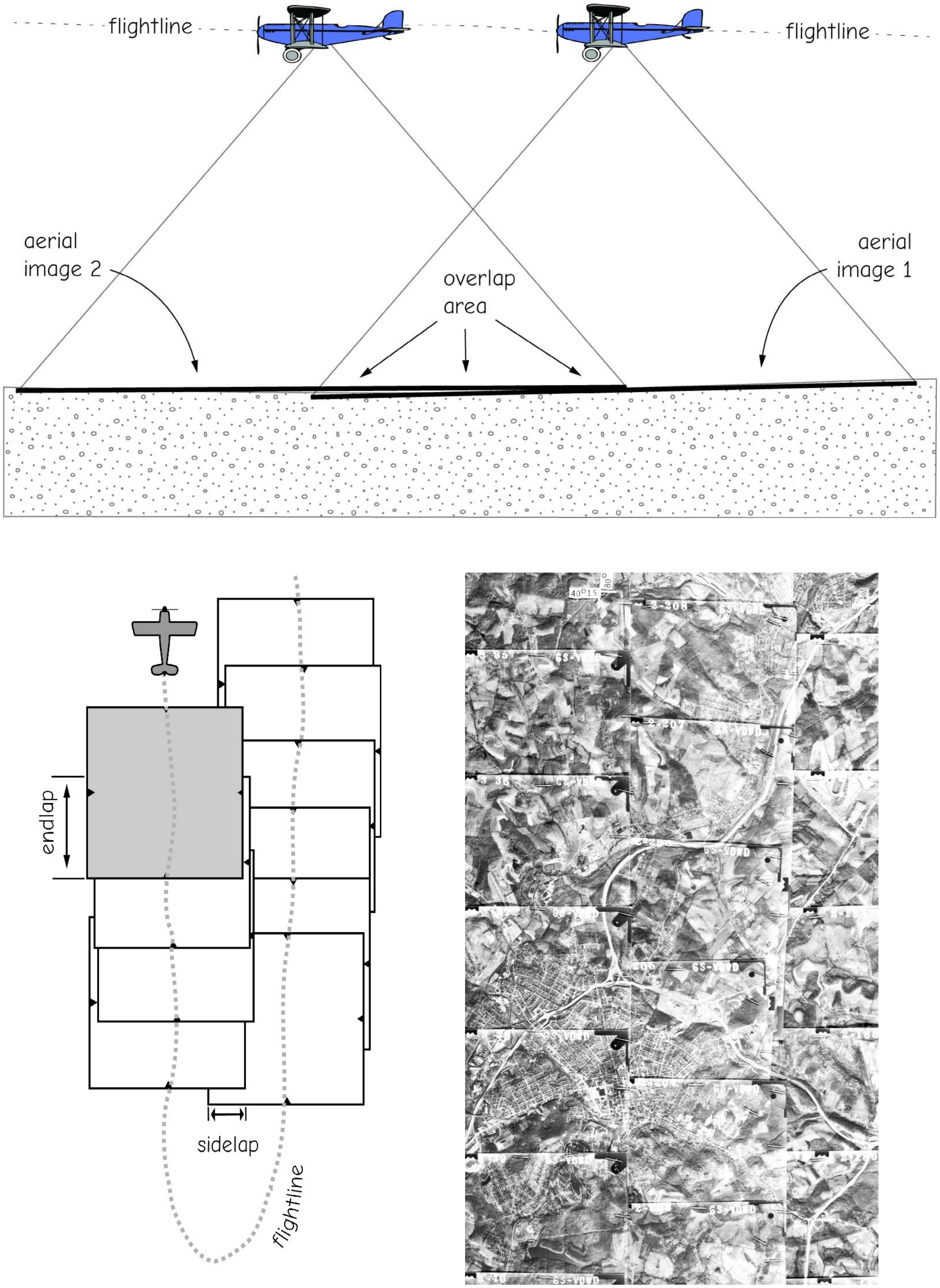


Figure 6-25: Aerial images often overlap to allow three-dimensional measurements and the correction of relief displacement. Sequential images along a flight path are taken at intervals such that there is an area of overlap, shown in the side view in the top figure. Flightlines may be spaced so that the edges of successive rows overlap, creating sidelap (bottom left). The grouped photos create a photomosaic (bottom right).

from the overlapping stereopair, we observe a set of relative spatial shifts in objects, and our brain may convert these to a perception of depth. When we have vertical aerial images, the distance from the camera to each point on the ground is determined primarily by the elevations at each point on the ground. We may observe parallax for each point and use this parallax to infer the relative elevation for every point.

Stereo viewing creates a three-dimensional stereomodel of terrain heights, with our left eye looking at the left photo and our right eye looking at the right photo. The three-dimensional stereomodel can be projected onto a flat surface and the image used to digitize a map. We may also interpret the relative terrain heights on this three-dimensional surface, and thereby estimate elevation wherever we have stereo coverage. We can use stereopairs to draw contour lines or mark spot heights. Before LiDAR, this was the most common method for determining elevation over areas larger than a few hundred hectares.

Stereomodels are visible in stereopairs due to *parallax*, a shift in relief displacement due to a shift in observer location. Figure 6-26 illustrates parallax. The block (closer to the viewing locations) appears to shift more

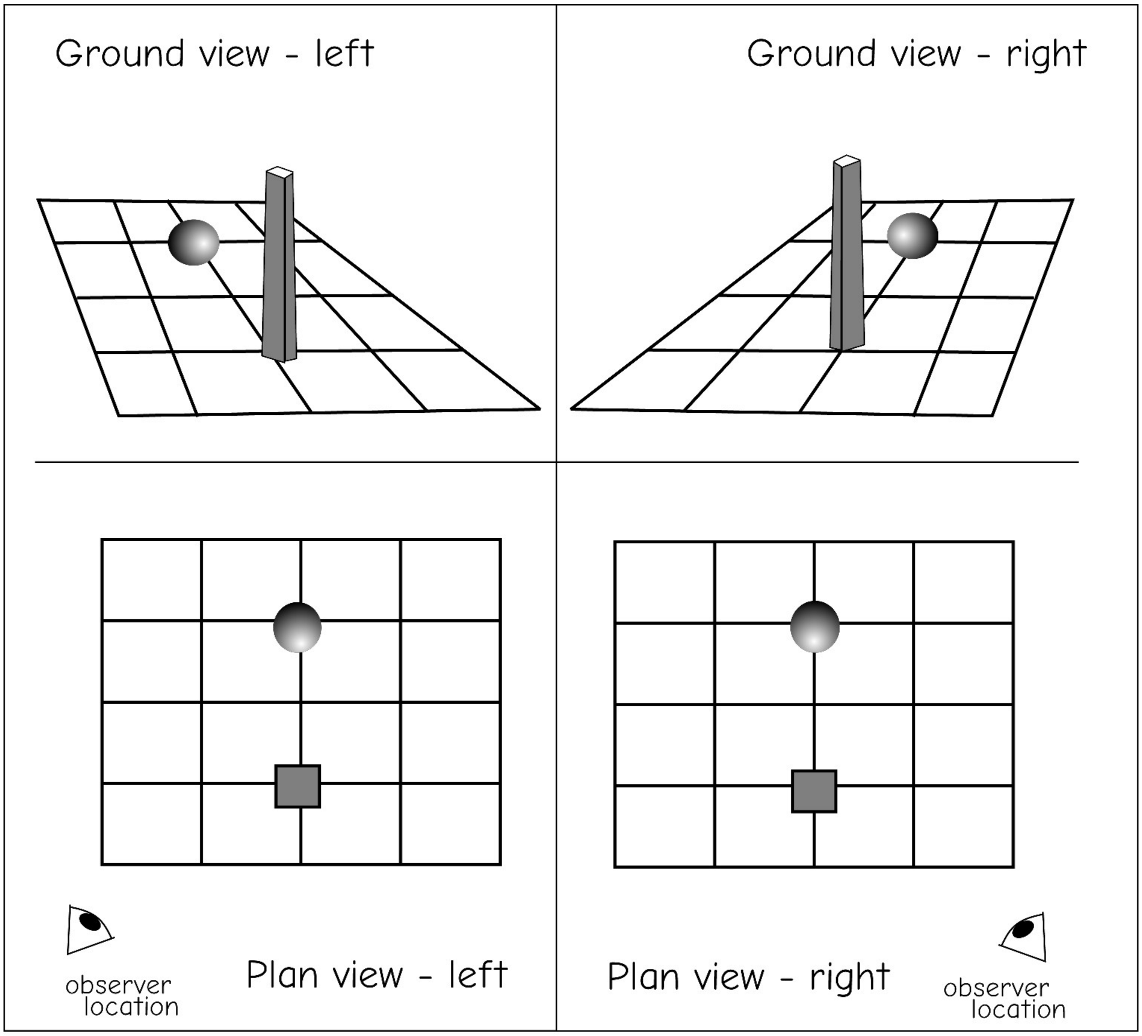


Figure 6-26: An illustration of parallax, the apparent relative shift in the position of objects with a shift in the viewer's position. Objects that are farther away (sphere, above) appear to shift more when a viewer changes position. This is the basis of stereo depth perception.

than the sphere when the viewing location is changed from the left to the right side of the objects. The displacement of any given point is different on the left vs. the right ground views because the relative viewing geometry is different. Points are shifted by different amounts, and the magnitude of the shift depends on the distance from the observer (or camera) to the objects. This shift in position with a shift in viewing location is by definition the parallax, and is the basis of depth perception.

Geometric Correction of Aerial Images

Due to the geometric distortions described above, it should be quite clear that uncorrected aerial images should not be used directly as a basis for spatial data collection under most circumstances. Points, lines, and area boundaries may not occur in their correct relative positions, so length and area measurements may be incorrect. These distortions are a complex mix of terrain and tilt effects, and will change the locations, angles, and shapes of features in the image and any derived data. Worse, when spatial data derived from uncorrected photographs are combined with other sources of geographic information, features may not occur at their correct locations. A river may fall on the wrong side of a road or a city may be located in a lake. Given all the positive characteristics of aerial images, how do we best use this rich source of information? Fortunately, photogrammetry provides the tools needed to remove geometric distortions from photographs. These corrections depend on two primary sets of measurements. First, the location of each image's *perspective center* or *focal center* must be known. This is approximately the location of the camera focal point at the time of imaging. It can be determined from precise GNSS, or deduced from ground measurements. Second, some direct or indirect measurement of terrain heights must be collected. These heights may be collected at a few points, and stereopairs used to estimate all other heights, or they may be determined from another source, for example, a previous survey, radar, or LiDAR systems described later in this chapter. Armed with perspective center and height measurements, we may correct our aerial images.

Geometric correction of aerial images involves calculating the distortion at each point, and shifting the image location to the correct orthographic position. Consider the tower in Figure 6-27. The bottom of the tower at B is imaged on the photograph at point b, and the top of the tower at point A is imaged on the photograph at point a. Point A will occur on top of point B on an orthographic map. If we consider the flat plane at the base of the tower as the datum, we can use simple geometry to calculate the displacement from a to b on the image. We'll call this displacement d, and go through an explanation of the geometry used to calculate the displacement.

Observe the two similar triangles in Figure 6-27, one defined by the points S-N-C, and one defined by the points a-n-C. These triangles are similar because the angles are equal, that is, the interior angle at n and N are both 90°, the triangles share the angle at C, and the interior angle at S equals the interior angle at a. C is the focal center of the camera lens, and may be considered the location through which all light passes. The film in a camera is placed behind the focal center; however, as in previous figures, the film is shown here in front of the focal center for

clarity. Note that the following ratios hold for the similar triangles:

$$D/P = h/H \quad (6.3)$$

and also

$$d/p = D/P \quad (6.4)$$

so

$$d/p = h/H \quad (6.5)$$

rearranging

$$d = p*h/H \quad (6.6)$$

where:

d = displacement distance

p = distance from the nadir point, n, on the vertical photo to the imaged point a

H = flying height

h = height of the imaged point

We usually know the flying height, and can measure the distance p. If we can get h, the height of the imaged point above the datum, then we can calculate the displacement. We might climb or survey the tower to measure its height, h, and then calculate the photo displacement by equation (6.6). Relief displacement for any elevated location may be calculated provided we know the height. Heights have long been calculated by measurements from stereopairs, but are increasingly measured using LiDAR, described

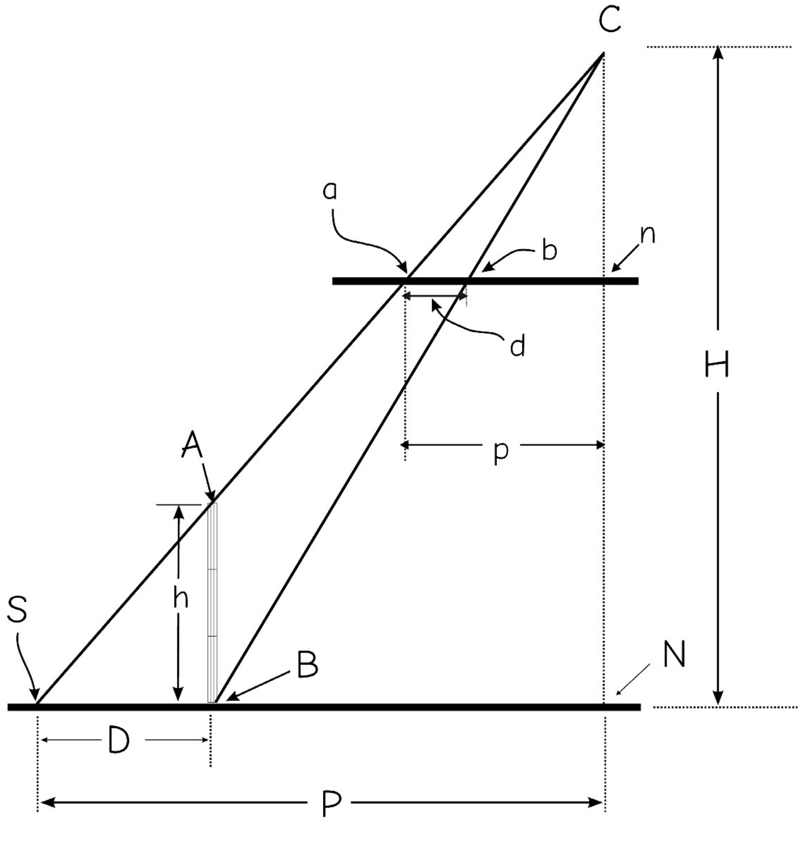


Figure 6-27: Relief displacement may be calculated based on geometric measurement. Similar triangles S-N-C and a-n-C relate heights and distances in the photograph and on the ground. We usually know flying height, H, and can measure d and p on the photograph.

later in this chapter. These heights and equations are used to adjust the positional distortion due to elevation, "moving" imaged points to an orthographic position.

Figure 6-28 illustrates the distortion in an image of a straight pipeline right-of-way, bent on the image by differences in height from valleys to rigdetops (left). Knowledge of image geometry allows us to correct the distortion (Figure 6-28 right).

Equation (6.6) applies to vertical aerial images. When photographs are tilted, the distortion geometry is much more complicated, as are the equations used to calculate tilt and elevation displacement. Equations may be derived that describe the three-dimensional projection from the terrain surface to the two-dimensional film plane. These equations and the methods for applying them are part of the science of photogram.

Typically, images are taken with a digital camera, or if taken with a film camera, the images are scanned. Measurements of image x and y are then determined relative to some image-specific coordinate system. These measurements are obtained from one or many images. Ground x, y, and z coordinates come from precise ground surveys.

A set of equations is written that relates image x and y coordinates to ground x, y, and z coordinates. The set of equations is solved, and the displacement calculated for each point on the image. The displacement may then be removed and an orthographic image or map produced. Distances, angles, and areas can be measured from the image. These orthographic images, also known as *orthophotographs* or *digital orthographic images*, have the positive attributes of photographs, with their rich detail and timely coverage, and some of the positive attributes of

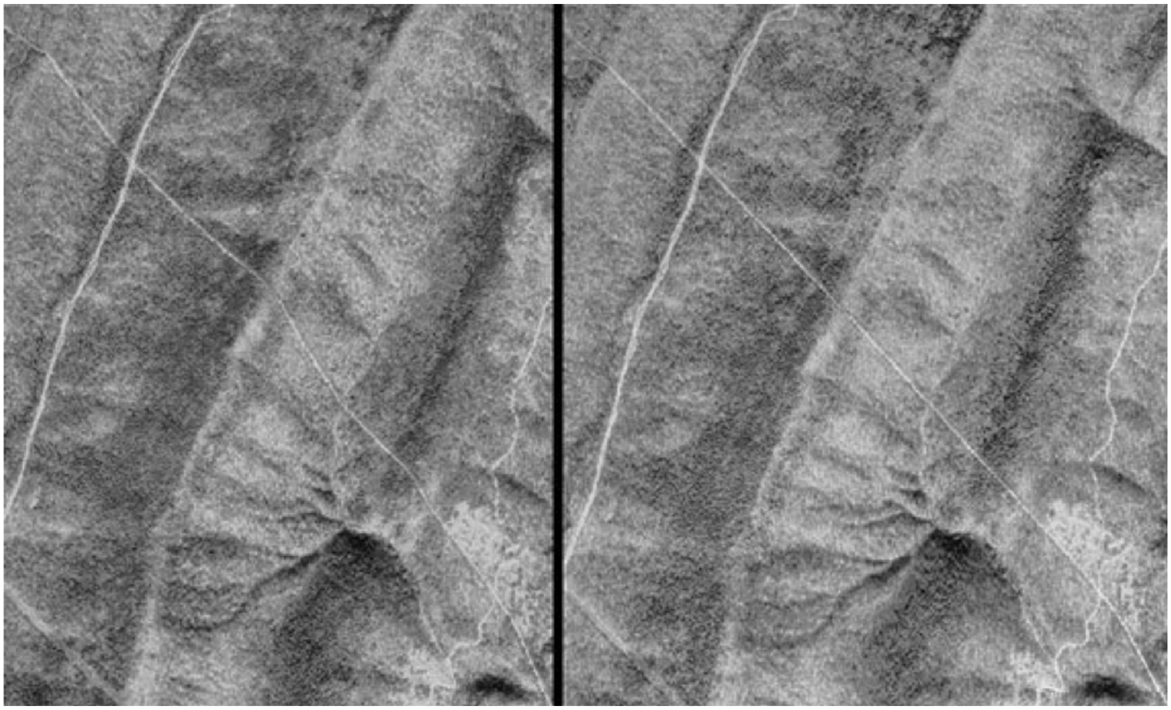


Figure 6-28: An example of distortion removal when creating an orthoimage. A nearly straight pipeline right-of-way spans uncorrected (left) and corrected (right) images, from the lower right to upper left in each image. The path appears bent in the image on the left as it alternately climbs ridges and descends into valleys. Using equations described in this section, these distortions may be removed, resulting in the orthographic image on the right, showing the nearly straight pipeline trajectory (courtesy USGS).

cartometric maps, such as uniform scale and true geometry.

Multiple images or image strips may be analyzed, corrected, and stitched together into a single mosaic. This process of developing photomodels of multiple images at once utilizes interrelated sets of equations to find a globally optimum set of corrections across all images.

Photo Interpretation

Aerial images are useful primarily because we may use them to identify the position and properties of interesting features. Once we have determined that the film and camera system meet our spatial accuracy and information requirements, we need to obtain the photographs, either from existing images in government or private archives, or new acquisitions in the field, and then interpret them. *Photo* (or image) *interpretation* is the process of converting images into information. Photo interpretation is a well-developed discipline, with many specialized techniques. We will provide a very brief description of the process. A more complete description may be found in several of the sources listed at the end of this chapter.

Interpreters use the size, shape, color, brightness, texture, and relative and absolute location of features to interpret images (Figure 6-29), typically digitizing directly on a screen displayed image, as described in Chapter 4. Differences in these diagnostic characteristics allow the interpreter to distinguish among features. In the figure, the polygon near the center of the image labeled Pa-C, a pasture, is noticeably smoother than the polygons surrounding it, and the polygon above it labeled As-Y1 shows a finer-grained texture, reflecting smaller tree crowns than the polygon labeled NH-M11 above it and to the left. Different vegetation types may show distinct color or texture variations, road types may be distinguished by width or the occurrence of a median strip, and building types may be defined by size or shape.

The proper use of all the diagnostic characteristics requires that the photo interpreter develop some familiarity with the features of interest. For example, it is difficult to distinguish the differences between many crop types until the interpreter has spent time in the field, photos in hand, comparing what appears on the photographs with what is found on the ground. This "ground truth" is invaluable in developing the local knowledge required for accurate image interpretation. When possible, ground visits should take place contemporaneously with the photographs.

Photo interpretation requires we establish a target set of categories for interpreted features. If we are mapping roads, we must decide what classes to use; for example, all roads will be categorized into one of these classes: unpaved, paved single lane, paved undivided multi lane, and paved divided multi lane. These categories must be inclusive, so that in our photos there must be no roads that are multi lane and unpaved. If there are roads that do not fit in our defined classes, we must fit them into an existing category, or we must create a category for them.

Photo interpretation also requires we establish a *minimum mapping unit*, or MMU. A minimum mapping unit defines the lower limit on what we consider significant, and usually defines the area, length, and/or width of the smallest important feature. The arrow in the lower right corner of Figure 6-29 points to a forest opening smaller than our minimum mapping unit for this example map. We may not be interested in open patches smaller than 0.5 ha, or road segments shorter than 50 m long. Although they may be visible on the image, features smaller than the minimum mapping unit are not delineated nor transferred into the digital data layer.

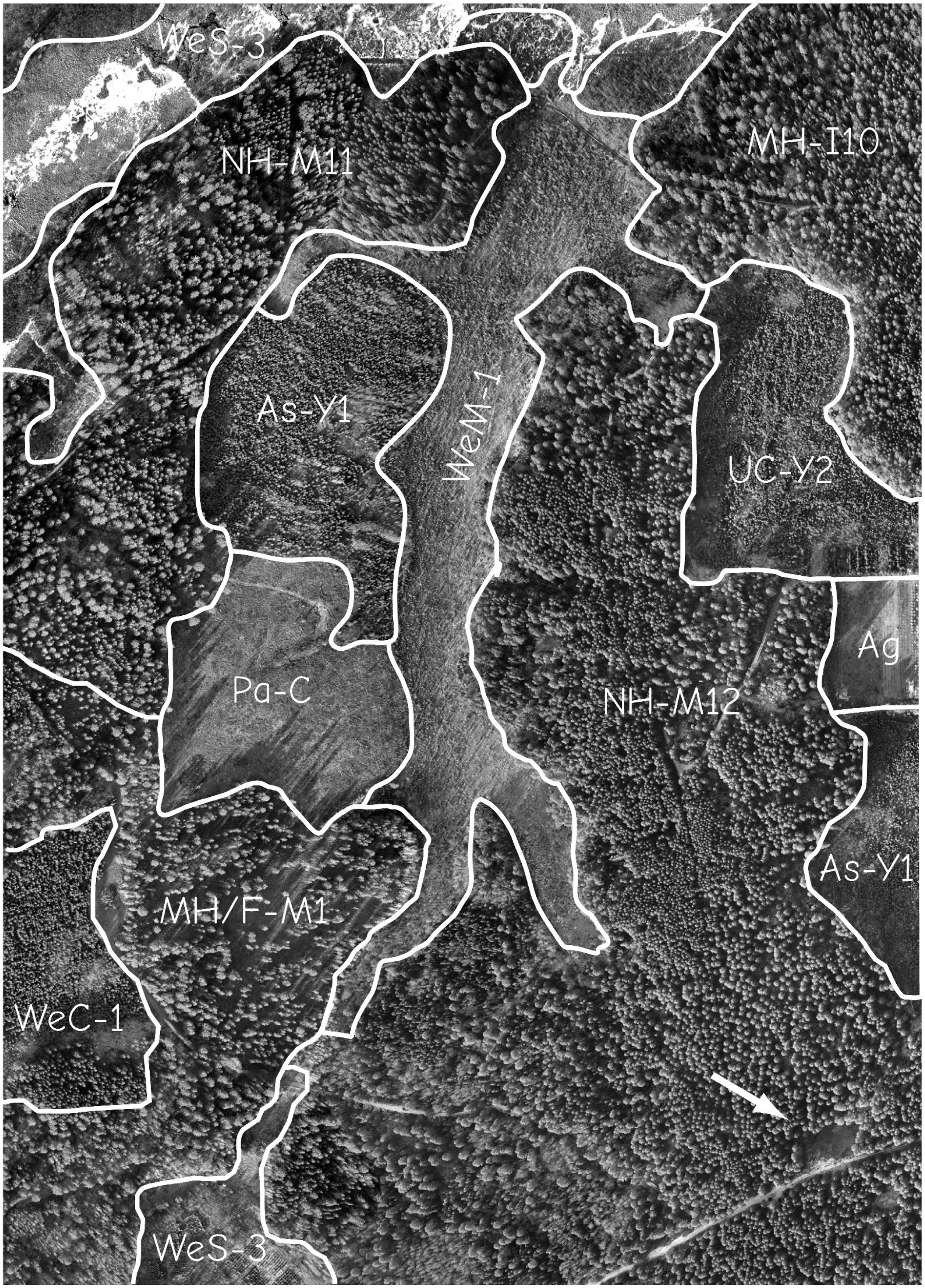


Figure 6-29: Photo interpretation is the process of identifying features on an image. Photo interpretation in support of GIS typically involves digitizing the points, lines, or polygons for categories of interest from a georeferenced digital or hardcopy image. In the example above, the boundaries between different vegetation types have been identified based on the tone and texture recorded in the image. The arrow at the lower right shows an "inclusion area", not delineated because it is smaller than the minimum mapping unit.

Images may need enhancement to improve feature identification. Common adjustments include band selection and display brightness, contrast, or image histogram modification. Bands typically keep a one-to-one correspondence with true color images, matching each output color to the respective input color. Selection is needed when more than three bands are collected, most commonly three visible plus near- or mid-infrared bands. The analyst must choose which bands to display and in which output color e.g., green, red, and infrared image layers to the blue, green, and red colors of the output display to yield a typical “false color” image. Some imaging systems collect additional mid-infrared bands, and different band combinations have proven best for specific features, e.g., a green, mid-infrared, infrared combination for some vegetation classifications.

Many images require modification of the display brightness, contrast, or histogram to optimize image interpretation (Figure 6-30). Most GIS software allow you to change the base color assigned to each layer, and manipulate the image histograms to enhance image display. You can optimize the various image histogram thresholds to best reveal the features you’re interested in identifying. Detailed descriptions of more sophisticated image enhancements can be found in most introductory remote sensing textbooks.

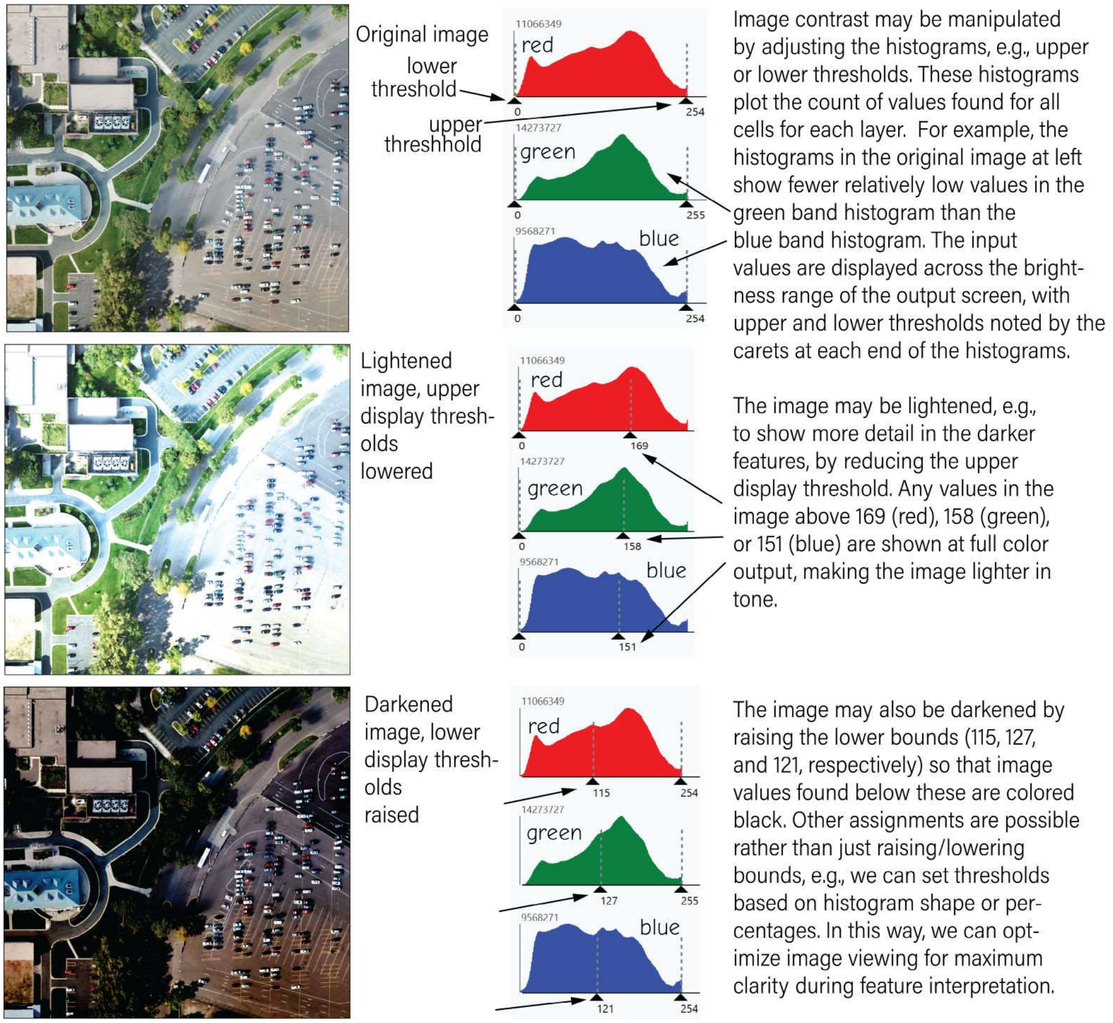


Figure 6-30: Examples of image bands and histogram manipulations.

Satellite Images

In previous sections we described the basic principles of remote sensing and the specifics of image collection and correction using aerial images. In many respects satellite images are similar to aerial images when used in a GIS. The primary motivation is to collect information regarding the location and characteristics of features. However, there are important differences between photographic and satellite-based scanning systems used for image collection, and these differences affect the characteristics and hence uses of satellite images.

Satellite scanners have several advantages relative to aerial imaging systems. Satellites offer a very high perspective, which significantly reduces terrain-caused distortion. Equation (6.6) shows the terrain displacement (d) on an image is inversely related to the flying height (H). Satellites have large values for H, typically 600 km (370 mi) or more above the Earth's surface, so relief displacements are correspondingly small. Because satellites are flying above the atmosphere, their pointing direction (attitude control) is very precise, and so they can be maintained in an almost perfect vertical orientation.

There are additional trade-offs in satellite vs. aerial platforms. Satellite images typically cover larger areas, so if the area of interest is small, costs may be needlessly high. Satellite images may require specialized image processing software. Acquisition of aerial images may be more flexible because a pilot can fly on short notice. Many aerial images have better effective resolution than satellite images. Finally, aerial images are often available at reduced costs from government sources. Many of these disadvantages of using satellite images diminish as more, higher-resolution, pointable scanners are placed in orbit.

Basic Principles of Satellite Image Scanners

Scanners operate by pointing the detectors at the area to be imaged. Each detector has an *instantaneous field of view*, or IFOV, that corresponds to the size of the area viewed by each detector (Figure 6-31). Although the IFOV may not be square and a raster cell typically is square, this IFOV may be thought of as approximately equal to the raster cell size for the acquired image.

The scanner builds a two-dimensional image of the surface by pointing a detector or detectors at each cell and recording the reflected energy. Data are typically collected in the across-track direction, perpendicular to the flight path of the satellite, and in the

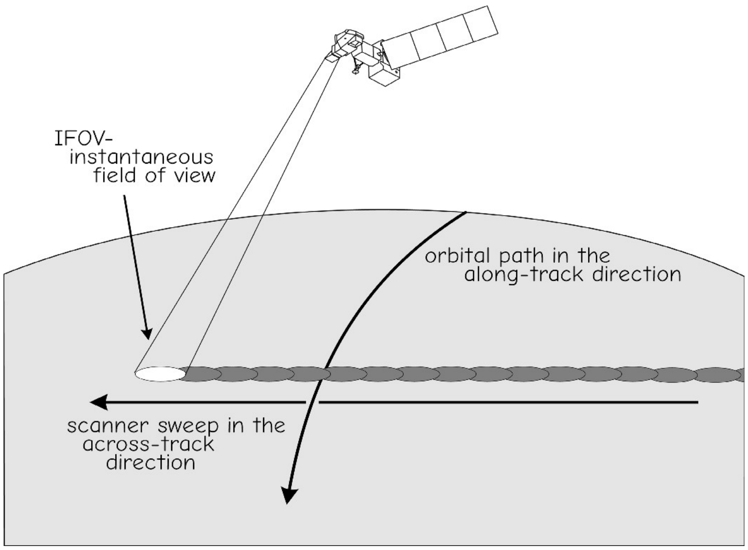


Figure 6-31: A spot scanning system. The scanner sweeps an instantaneous field of view (IFOV) in an across-track direction to record a multispectral response. Subsequent sweeps in an along-track direction are captured as the satellite moves forward along the orbital path. Line- or grid-capture systems are similar, recording multiple IFOVs simultaneously.

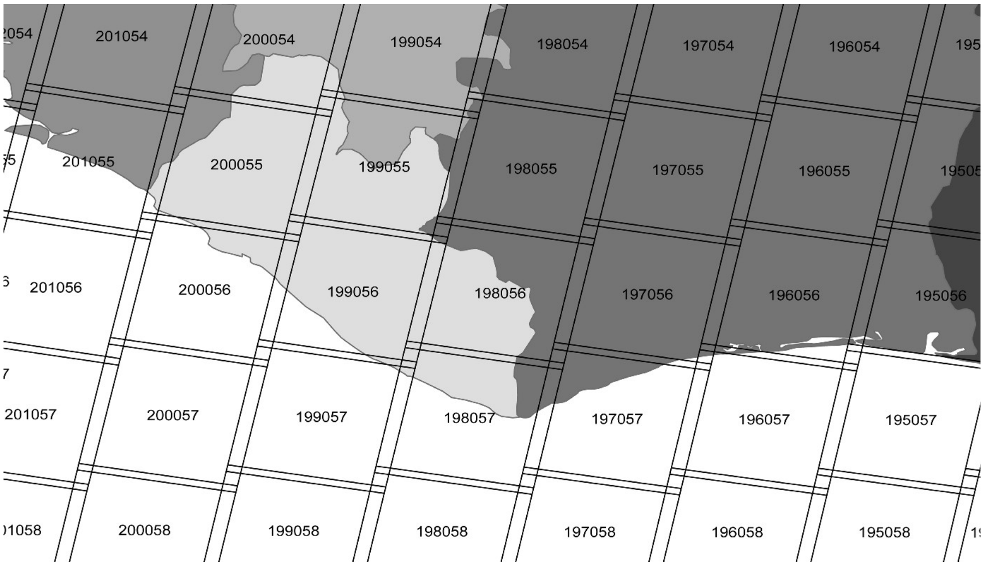


Figure 6-32: A portion of the path and row layout for the Landsat satellite systems. Each slightly overlapping, labeled rectangle corresponds to a satellite image footprint.

along-track direction, parallel to the direction of travel (Figure 6-31). Several scanner designs achieve this across- and along-track scanning. Some older designs use a *spot detector* and a system of mirrors and lenses to sweep the spot across track. The forward motion of the satellite positions the scanner for the next swath in the along-track direction. Other designs have a *linear array* of detectors – a line of detectors in the across-track direction. The across-track line is sampled at once, and the forward motion of the satellite positions the array for the next line in the along-track direction. Finally, a *two-dimensional array* may be used, consisting of a rectangular array of detectors. Reflectance is collected in a patch in both the across-track and the along-track directions.

A remote sensing satellite also contains a number of other subsystems to support image data collection. A power supply is required, typically consisting of solar panels and batteries. Precise altitude and orbital control are needed, so satellites carry navigation and positioning subsystems. Sensors evaluate satellite position and pointing direction, and thrusters and other control components orient the satellite. There is a data storage subsystem, and a communications subsystem for transmitting data back to Earth and for receiving control and other information. All of these activities are coordinated by an onboard computing system.

Several remote sensing satellite systems have been built, and data have been available for land surface applications since the early 1970s. The detail, frequency, and quality of satellite images have been improving steadily, and there are several satellite remote sensing systems currently in operation.

Satellite data are often nominally collected in a path/row system. A set of approximately north-south paths are designated, with approximately east-west rows identified across the paths. Satellite scene location may then be specified by a path/row number (Figure 6-32). Satellite data may also be ordered for customized areas, depending on the flexibility of the acquisition system.

Because most satellites are in near-polar orbits, images overlap most near the poles. Adjacent images typically overlap a small amount near the equator. The inclined orbits are often sun synchronous, meaning the satellite passes overhead at approximately the same local time.

High-Resolution Satellite Systems

There is a large and growing number of high-resolution satellite systems, here rather arbitrarily defined as those with a resolution finer than 4 m. This is the resolution long available on the largest-scale aerial photographs, and used for fine-scale mapping of detailed features such as sidewalks, houses, roads, individual trees, and small-area landscape change. Commercial systems providing 30 cm resolution are in operation (Figure 6-33), with higher-resolution systems in the offing. This detail blurs the distinction between satellite and photo-based images.

Images from high-resolution satellite systems may provide a suitable source for spatial data in a number of settings. These images provide substantial detail of manmade and natural features, and match the spatial resolution and detail of high-accuracy GNSS receivers. They are typically required by cities and businesses for fine-scale asset management, for example, in urban tree inventories, construction monitoring, or storm damage assessment. Nearly all the systems have pointable optics or satellite orientation control, resulting in short revisit times, on the order of one to a few days.

Spectral range, price, availability, reliability, flexibility, and ease of use may become more important factors in selecting between aerial images and high-resolution satellite images. Satellite data are attractive when collecting data for larger areas, or where it is unwise or unsafe to operate aircraft, or because data for large areas may be geometrically corrected for less cost and time. Aerial images may be preferred when resolutions of a few centimeters are needed, or for smaller areas, under narrower acquisition windows, or with instrument clusters not possible from space. Aerial images will not be completely replaced by satellites, but they may well be pushed towards the finest resolutions and county-sized or smaller collections.

Figure 6-33: A 0.3 m resolution image of the Kalgoolie Mine in Western Australia, demonstrating the detail available from the highest-resolution satellite imaging systems (courtesy Digital Globe.).

As of early 2019, there are several operational satellite systems capable of global image acquisition at 1 m resolution or better, including WorldView, GeoEye, Pleiades, Kompsat, and SPOT. These satellites and related systems are predominantly commercial enterprises, funded and operated by businesses. There are several recently decommissioned high-resolution systems for which archive data are available and still useful, including the Ikonos satellite that operated from 1999 through early 2015, and the Quickbird system, operational from 2001 through early 2015.

The WorldView-3 and WorldView-4 satellites currently provide the highest resolution available on a global basis, with a maximum resolution of 31 cm provided as panchromatic images, with addition of eight bands at a 1.24 m resolution, eight shortwave infrared bands at 3.7 m resolution for haze and smoke penetration, and 12 bands at 30 m resolution. Images are collected in a 13.1 km swath width at nadir, and due to satellite reorientation has a revisit time of less than a day, with effective global coverage on a 3-day basis. Off-nadir resolution is poorer than 30 cm, but each satellite can image the entire globe at better than 50 cm resolution in a 4.5-day period. Images are collected at approximately 10:30 a.m. local time, a common characteristic of these polar orbiting, sun-synchronous systems.

WorldView-1 and -2 preceded the current satellites, and are still collecting data. WorldView-1 provides 0.5 m panchromatic images, while the WorldView-2 provides 0.46 m panchromatic and multispectral images at 1.8 m (Figure 6-34). Data are collected as often as a 1.7-day return interval when providing a 1 m resolution, and 6 days with a 0.5 m resolution. Images have a swath width of approximately 17 km.

GeoEye-1 was launched in mid-2008 into a sun-synchronous orbit with a local collection time near 10:30 a.m. Image resolution has changed with time, but it currently collects panchromatic images with a 46 cm resolution and multispectral images span-

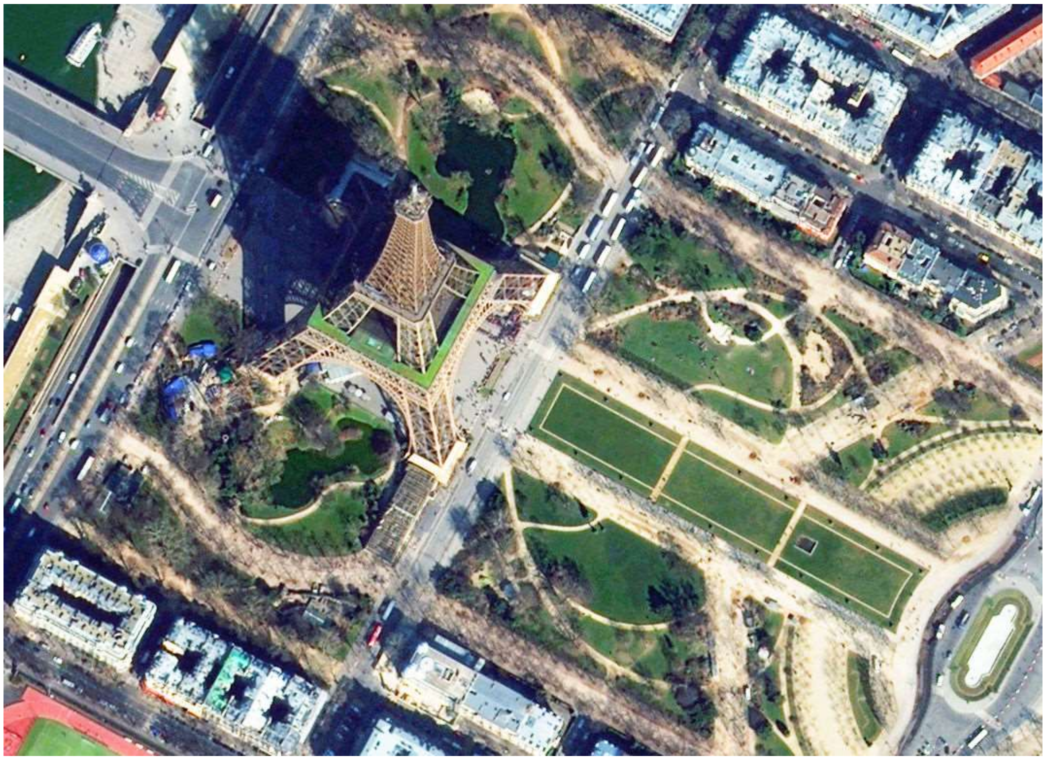


Figure 6-34: A 0.5 m resolution image from the WorldView-2 satellite (courtesy DigitalGlobe).

Figure 6-35: A SPOT-6 image of Bora-Bora, demonstrating the high resolution over a relatively large area available from the system (courtesy SPOT Image Corp.).

ning the blue through near-infrared portions of the spectrum at a 1.84 m resolution. There is a nominal 15.2 km scan width at nadir, and off-nadir imaging allows return intervals of as short as 2 days. Although the system has a 7 year design life, it may well operate much longer, given satellite imaging systems have often functioned to double their designed interval.

Pleiades-1 and -2 were launched in late 2011 and 2012 by a European consortium, with a five-year design life. The system provides 50 cm panchromatic and 2 m multispectral data in blue through near-infrared bands, with a 20 km swath width at nadir. Sun-synchronous orbital planes are offset by 180°, with a pointable satellite, allowing daily revisits by the constellation.

Another set of high-resolution images come from the *Système Pour l'Observation de la Terre* (SPOT), versions SPOT-6 and SPOT-7. These are an evolution of a set of mid-resolution satellites, SPOT-1 through -5, described in the next section. The high-resolution satellites carry a 1.5 m panchromatic and 6 m resolution multispectral scanner, the latter with four bands spanning the visible through near-infrared spectrum (Figure 6-35). SPOT has a 60 km swath width at nadir. Note that this larger swath width provides 15 to 40 times the area coverage of the highest-resolution satellites, and illustrates a more general trade-off between satellite image resolution and the area covered by each image. The set of SPOT satellites has a daily revisit capability, completely covering the Earth's landmasses every two months.

The Dove satellite cluster by Planet Inc. carries this notion of a constellation of small, inexpensive, high-resolution satellites further, with a fleet of more than 130 satellites, approximately the size of a rural

route mailbox, inexpensively deployed in clusters. First satellites were launched in 2013, with a group deployment of 28 satellites in 2014, and the full constellation during 2015. There were 150 operational Dove satellites in early 2019.

Images have 3 m resolution at nadir, although may reach 5 m in some configurations. The constellation provides daily revisit times globally, with higher frequency revisits depending on orbital patterns, location on Earth, and satellite tasking. Images are stitched together for complete global coverage, updating a global mosaic on a daily basis.

A number of high-resolution satellite imaging systems have a local focus. The KOMPSAT-2 satellite is designed to collect data primarily over eastern Asia, and provides 1 m resolution panchromatic and 4 m multispectral data. The Cartosat-2 satellite, launched in 2007, provides 0.9 m resolution panchromatic data, primarily focused on south Asia.

There are additional systems in operation. Four DMC3 satellites have been launched by Surrey Satellite Technology Ltd., with 1 m panchromatic and 4 m multispectral images in a 24 km swath width. Launches span 2013 through 2018 with a seven year satellite design life, although in practice most satellites last longer.

Skysat-1 was launched in late 2013 and is notable for providing high-resolution images based on a small satellite, low cost approach. This approach may lead to both a larger constellation of satellites with more frequent revisits, and lower-priced images. Currently, there is a planned constellation of 15 satellites, with launches scheduled at least through 2018. Skysat provides a 90 cm single-band panchromatic mode and four, 2 m resolution multispectral bands, the latter in the blue through near-infrared region. There is a nominal 8 km swath width at nadir of the multispectral data, and 2 km for panchromatic images. Repeat intervals will depend on the number of satellites deployed.

Gaofen-1 through -4 make up a Chinese high-resolution satellite system designed for disaster prevention and relief, climate change monitoring, and agricultural and natural resource monitoring and support. Satellites carry differing instruments, but among them are 1 and 2 meter resolution panchromatic imagers.

Mid-Resolution Satellite Systems

There are several mid-resolution satellite systems, here defined as those providing images with resolutions from 5 m to less than 100 m. These are most often used for medium- to broad-area analyses, for example, landcover mapping at county, regional, or national extents, or large-area wildfire or flooding management. Individual image collections are generally several tens to hundreds of kilometers on a side, and revisit times from a few days to a few weeks.

SPOT

SPOT is one of the longest running, uninterrupted satellite imaging systems. The French Government led the development of SPOT, with SPOT-1 launched in early 1986. There were four additional SPOT mid-resolution satellites, labeled two through five, all since decomissioned. The two operating high/mid-resolution upgrades, SPOT-6 and -7 were described in the previous section because they offer 2 m resolution panchromatic images. They are included here because they maintain continuity with previous satellites by collecting similar multispectral bands, at a 6 m resolution.

There is a large archive of early SPOT images, useful for time series analysis and change detection. These mid-resolution satellites offered panchromatic and high-resolution visible (HRV) modes. Panchromatic resolution was between 2.5 and 10 m, with visible through mid-infrared bands at 10 to 20 m. This combination provides high resolution over large areas (Figure 6-36), and SPOT data are routinely used in a number of

Figure 6-36: An example of an image from the SPOT satellite system. The active Mexican volcano Popocatepetl is visible at the image center. This image demonstrates the broad area coverage and fine detail available from the SPOT system (courtesy SPOT Image Corp.).

resource management, urban planning, and other applications.

The SPOT scanners were among the first to have optics pointable to areas up to 27° to either side of the satellite path. This reduces revisit time to between one and five days, and allows the collection of satellite stereopairs suitable for elevation mapping.

Landsat

The Landsat-8 satellite is the latest in the longest running series of mid-resolution imaging satellites. Landsat-8 collects a 15 m resolution panchromatic band, 8 multispectral bands at 30 m in the visible, near-infrared, and mid-infrared portions of the spectrum, and two bands in the thermal infrared range with a 100 m resolution. The system has a 185 km swath width at nadir and a repeat interval of 16 days.

Landsat-8 uses an instrument called the *Operational Land Imager* (OLI) to collect non-thermal bands. The specific bands used were selected to be compatible with previous Landsat missions, and to improve cloud detection and aerosol/atmospheric haze analysis while surface mapping. The OLI also increases the bit depth, or data width from 8 to 12 bits, giving a broader and more sensitive response, and clearer, more detailed images.

Because Landsat was the first Earth-observing satellite system and it has operated nearly continuously since 1972, there is an image repository spanning five decades. The majority of these images (Figure 6-37)

are available free of charge to anyone with an internet connection, allowing long-term monitoring and analysis. Landsat-8 images are processed and added to this archive, typically within a few days of collection, resulting in an inexpensive source of broad-scale images. This long time series is particularly appropriate for change analysis, provided the differences between legacy and new data resolutions and formats are addressed.

Previous Landsat satellites have carried three primary imaging scanners. The *Multispectral Scanner* (or MSS) was the first satellite-based land scanner, launched in 1972, and it has been carried on board Landsat satellites 1 through 5. The original MSS sensed in four spectral bands, at an 80 m resolution: a green, a red, and two infrared bands.

Starting in 1984, Landsat satellites also carried the Thematic Mapper (TM) or Enhanced Thematic Mapper (ETM+), an improvement over the MSS. TM data contain seven spectral bands (three visible, a near-infrared, two mid-infrared, and a thermal band), and a 28.5 m grid-cell resolution for the first six bands. The ETM+ added a 15 m resolution panchromatic band covering the visible wavelengths. The satellites have had a 16 to 18 day return interval.

Landsat is used in many projects worldwide because of the breadth of radiometric bands, the large scan area for individual images, the long data record, and no-cost data. Landsat is the basis of many statewide and national land cover mapping projects, and it has been used to assess water quality in lake and coastal areas. Landsat is particularly appropriate for change detection, and much work has established methods for radiometric correction through time and across sensors, so that the time series of images may be used to map urban growth, vegetation change, and trajectories in water quality.

Figure 6-37: An example of a Landsat-5 image, showing the Mississippi River Delta. Mid-resolution satellites are particularly appropriate for regional or other large-area analysis (courtesy NASA).

Sentinel

The Sentinel system, launched by the European Space Agency (ESA), is comprised of six missions, including atmospheric, oceanic, and land resources (Figure 6-38). Each mission consists of two satellites, typically in offset orbits to provide maximum coverage and frequent repeat observations.

Sentinel-2 frequently contributes to GIS, as it provides land surface measurements including landcover classification, vegetation type, structure, and health, and snow cover and hydrology. It images in 12 spectral bands, from below blue wavelengths through shortwave infrared. Cell resolution varies by band from 10 to 60 meters, with a full complement of visible and near-infrared bands provided at 10 m resolution. Images are in a 290 km (180 mi) cross width, and variable length strips. Data are recorded in 12 bits, providing 16 times the radiometric resolution of 8 bit data. The two satellites together provide a five-day return frequency and global coverage, high among mid-resolution satellites. Data are free after registration with an ESA distribution portal.

Resourcesat

The Indian Space Research Organization has launched a number of satellites designed for Earth observation, including the Cartosat series, previously described, and the IRS series, beginning with the IRS-1A in 1988. Early satellites were largely experimental and images not widely distributed, or classified, but two, Resourcesat-1 and Resourcesat-2, provide high-quality, large-area, moderate resolution data over much of the globe (Figure 6-39).

The Resourcesats have carried three scanners, LISS-IV with a 5.8m resolution,

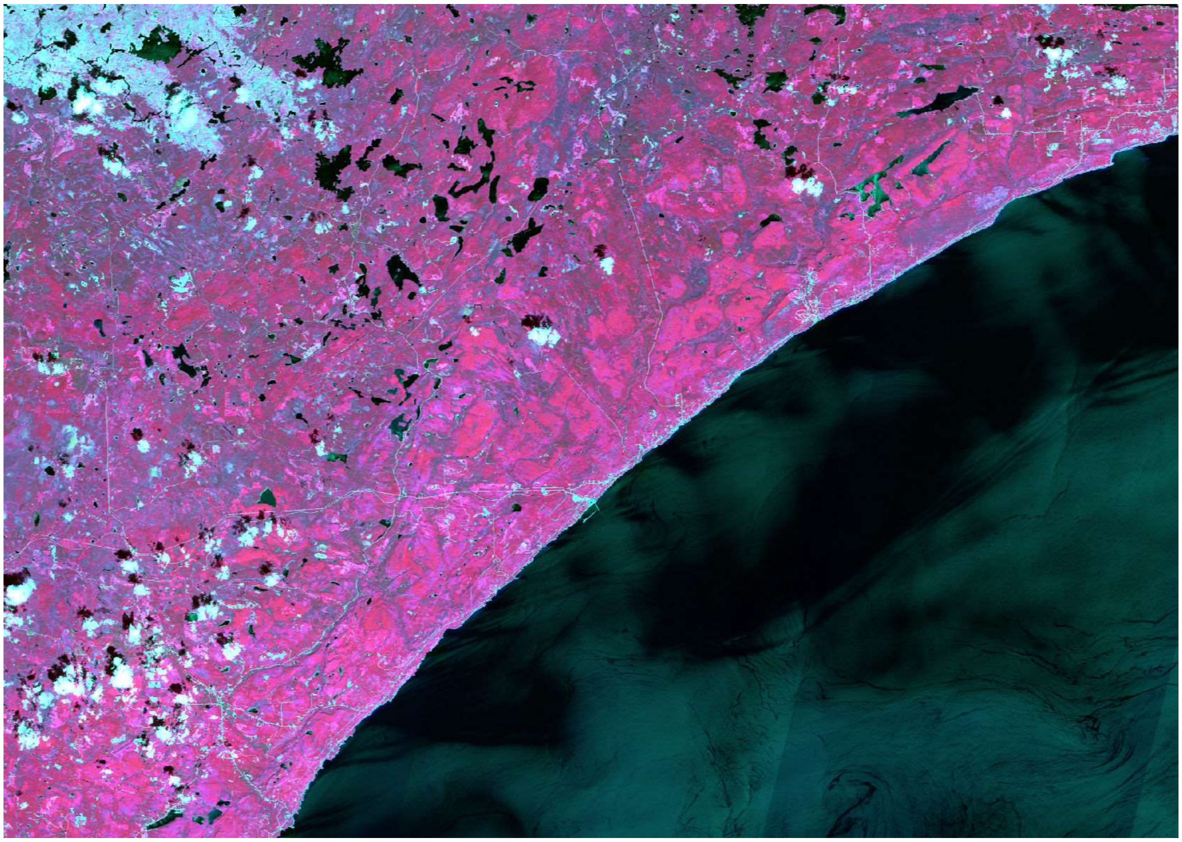


Figure 6-38: The Sentinel satellite system provides global mid-resolution data across a broad spectral range.

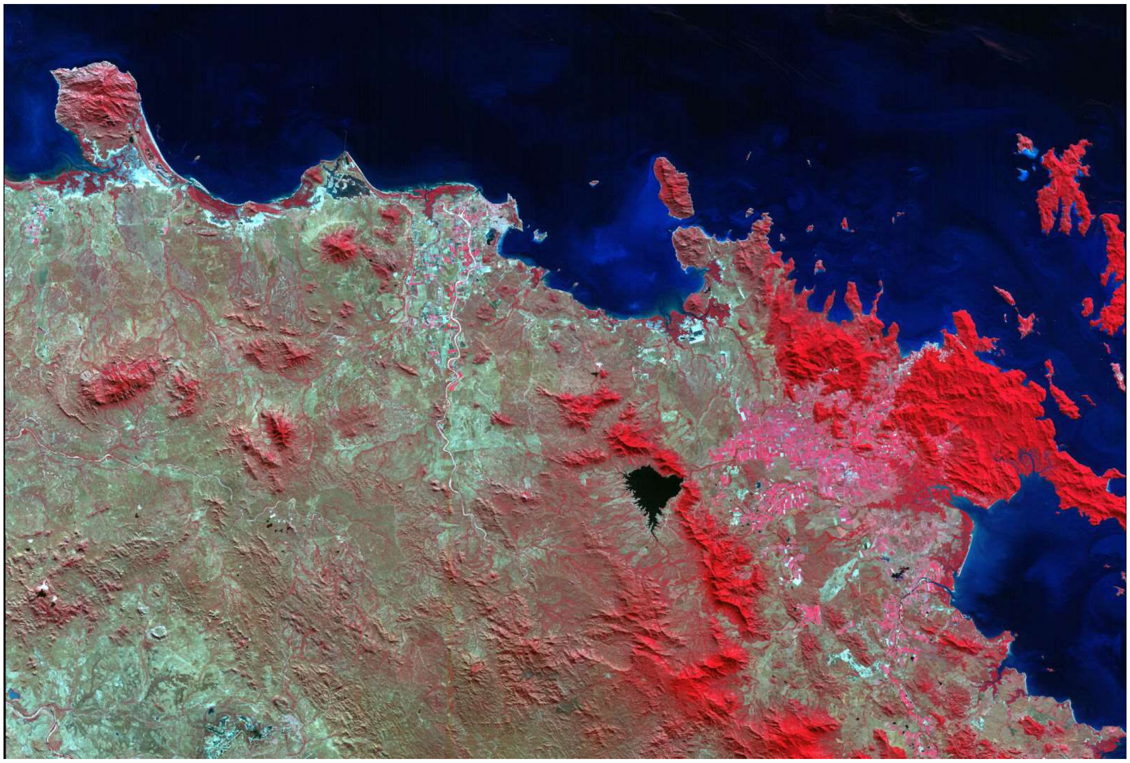


Figure 6-39: A Resourcesat LISS-III visible/near-infrared composite of a coastal area. Resourcesat provides medium-resolution data over large areas at reasonable costs.

the LISS-III with a 23.5m resolution, and AWiFS with a 56 m resolution. Swath width increases from 70 km through 141 km to 740 km for the three instruments, with a 5-day repeat cycle for the AWiFS sensor. The AWiFS is most commonly used outside of India, and provides blue, green, red, and near-infrared sensing bands, with 10-bit data. These images are often used for regional to national analyses because of their large image size and medium resolution, for example, by the U.S. National Agricultural Statistical Service for annual crop inventories in the United States.

RapidEye

RapidEye is a five-satellite constellation that provides images with an up to 5 m resolution, in five bands spanning the blue through near-infrared spectrum. Swath width is 77 km at nadir, with a 5.5 day repeat interval for nadir collections. A single-day repeat interval is possible for off-nadir viewing, but as with all tilted collections, at a reduced resolution. Satellites were launched simultaneously in 2008, with a seven year design life. Successor satellites are currently planned.

RapidEye is perhaps characteristic of a new era of high- and mid-resolution remote sensing systems. It was developed and deployed by a private entity, and ownership has changed during the system life. Although the most current technology and agile acquisition and delivery are supported, continuity of acquisition for the specific sensors is less certain. These systems may provide advantages for once-off or short-span applications, as with specific disaster assessments or annual crop mapping, but may provide disadvantages for long-term monitoring or change detection, for example, decades-long land cover change.

Coarse-Resolution, Global Satellite Systems

There are currently two widely used, coarse-resolution sensors: the Moderate Resolution Imaging Sensor, or MODIS, and the Visible Infrared Imaging Radiometer Suite, or VIIRS. MODIS is a NASA research system that collects data at a range of resolutions and wavebands, from visible through thermal infrared bands. Resolutions depend on bands and vary from 250 m to 1 km, and it has a repeat frequency of every one to two days for the entire Earth's surface when images are sampled at the 1 km resolution. Thirty-six bands are collected when operated in the 1 km mode, ranging from 0.4 μm to 14.4 μm. Only two bands are collected at the 250 m resolution, one each in the red and infrared portions of the light spectrum. These are somewhat unique in that the resolution is finer than the 1 km resolution of previous daily coverage satellite data, but substantially coarser than Landsat, SPOT, and moderate-resolution satellites. Large area coverage is possible at an intermediate level of detail when using MODIS 250 m data (Figure 6-40).

The MODIS satellites were launched in late 1999/mid 2000, with a six-year design life, so they have already outlived their mission, although they continue to function.

VIIRS is a successor instrument to MODIS, created to collect data for weather, ocean, and land surface analysis (Figure 6-41). It collects 9 visible near-infrared bands plus a day/night band, 8 mid-infrared bands, and 4 long-wave infrared bands. VIIRS collects data at both 375 and 750 m resolutions, and 3,040 km wide swath width, providing global coverage on a daily basis. It maintains continuity in some of the MODIS bands and products.

Figure 6-40: A MODIS 250 m resolution image of northern Italy and Switzerland. The snow-covered Alps cross through the center of this image, north of the Po River valley in Italy. Small cumulus clouds are visible, as is turbidity in the Mediterranean Sea and variation in land cover (courtesy NASA).

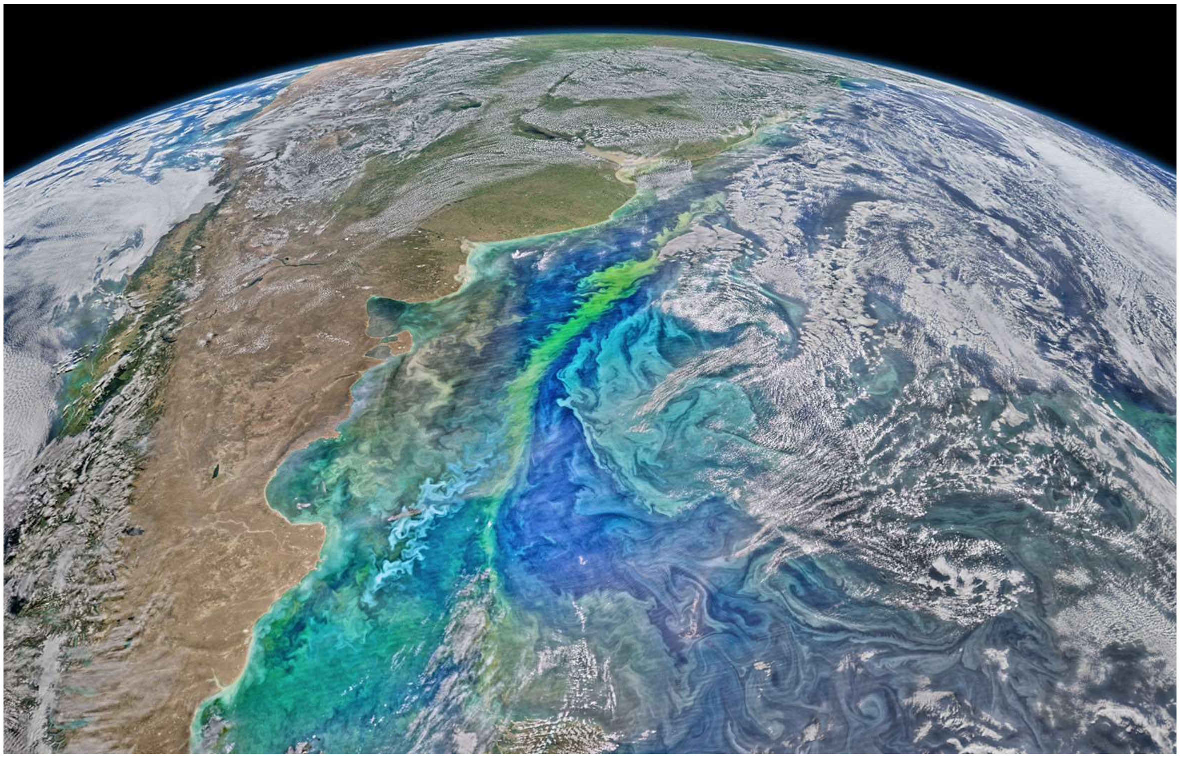


Figure 6-41: An image illustrating the broad area coverage and information content available from the VIIRS system. High concentrations of phytoplankton are visible along the east coast of South America, as well as arid and productive areas on land, and cloud and snow cover.

VIIRS is a substantial improvement over previous coarse-resolution satellites in several ways. Bands are tailored to provide enhanced information in specific windows, with improved vegetation, ocean color and productivity, land and ocean temperature, and cloud, fire, smoke, and sea ice detection. VIIRS data are freely available via U.S. NOAA portals.

Other Systems

There are several other airborne and satellite remote sensing systems that are operational or under development. Although some are quite specialized, each may serve as an important source of data. Some may introduce entirely new technologies, while others replace or provide incremental upgrades to existing systems. Space prevents our offering more than a brief description of these satellite systems here.

Passive optical systems: there are several remote sensing systems that are based on reflected incident radiation. These include the IRS system deployed by the Indian government, with a 5 m panchromatic band and a five-day revisit interval, three 24 m bands that span the green through near-infrared portion of the spectrum, and one 70 m band in mid-infrared portion of the spectrum.

A number of radar-based satellite systems have been used as a source of spatial data for GIS. Radar wavelengths are much longer than optical remote sensing systems, from approximately one to tens of centimeters, and may be used day or night, through most weather conditions. Radar images are panchromatic, because they provide information on the strength of the reflected energy at one wavelength. Radar systems have been successfully used for topographic mapping and some landcover mapping, particularly when large differences in surface texture occur, such as between water and land, or forest and recently clearcut areas. Operational systems include the ERS-1,

operated by the European Space Agency; the JERS-1, by the National Space Development Agency of Japan; and the Radarsat system, developed and managed by the Canadian Space Agency.

Satellite Images in GIS

Satellite images have two primary uses in GIS. First, satellite images are often used to create or update landcover data layers. Satellite images are particularly appropriate for landcover classification by virtue of their uniform data collection over large areas. Landcover classes often correspond to specific combinations of spectral reflectance values. For example, forests often exhibit a distinct spectral signature that distinguishes them from other landcover classes (Figure 6-42).

Satellite image classification involves identifying the reflectance patterns associated with each landcover class, and then applying this knowledge to classify all areas of a satellite image. Many techniques have been developed to facilitate landcover mapping using satellite data, as well as techniques for testing the classification accuracy of these landcover data. Regional and statewide classifications are commonly performed, and these data are key inputs in a number of resource planning and management analyses using GIS.

Satellite images are also used to detect and monitor change. The extent and intensity of disasters such as flooding, fires, or hurricane damage may be determined using satellite images. Urbanization, forest cutting, agricultural change, or other changes in land use or condition have all been successfully

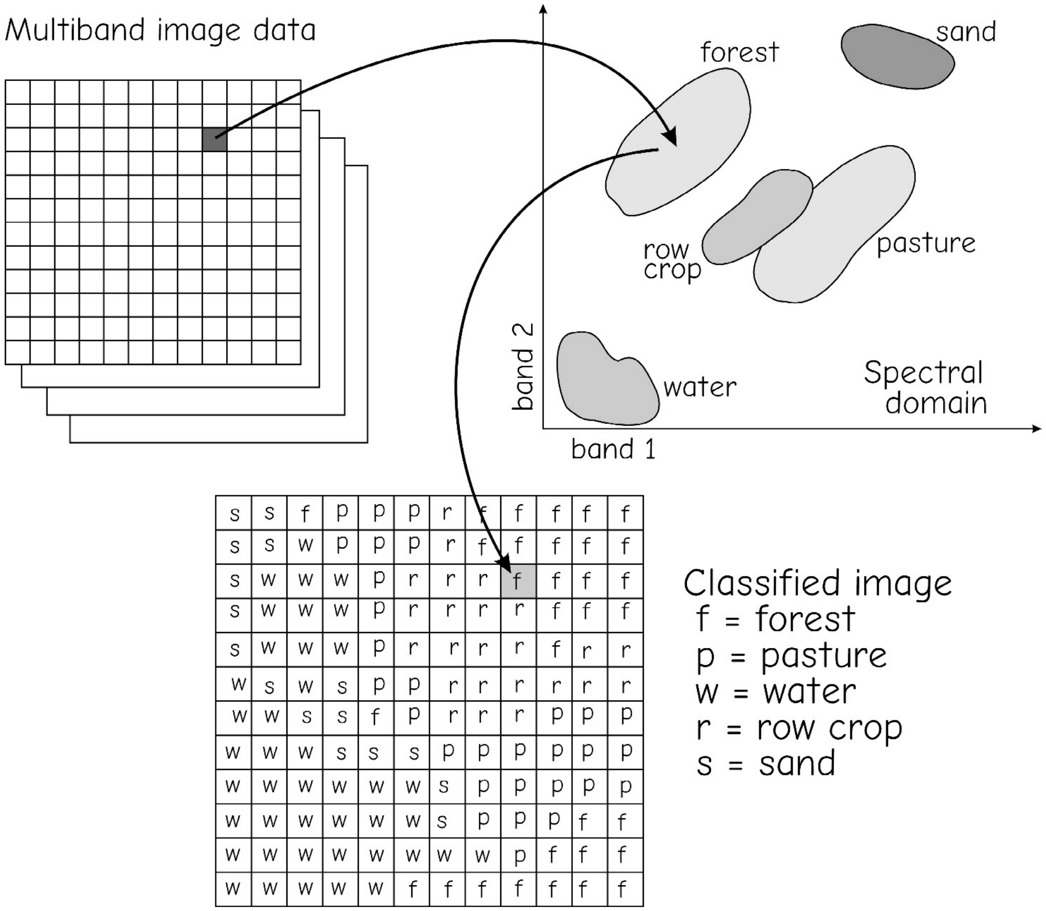


Figure 6-42: Landcover and land use classification is a common application of satellite images. The spectral reflectance patterns of each cover type are used to assign a unique landcover class to each cell. These data may then be imported into a GIS as a raster data layer.

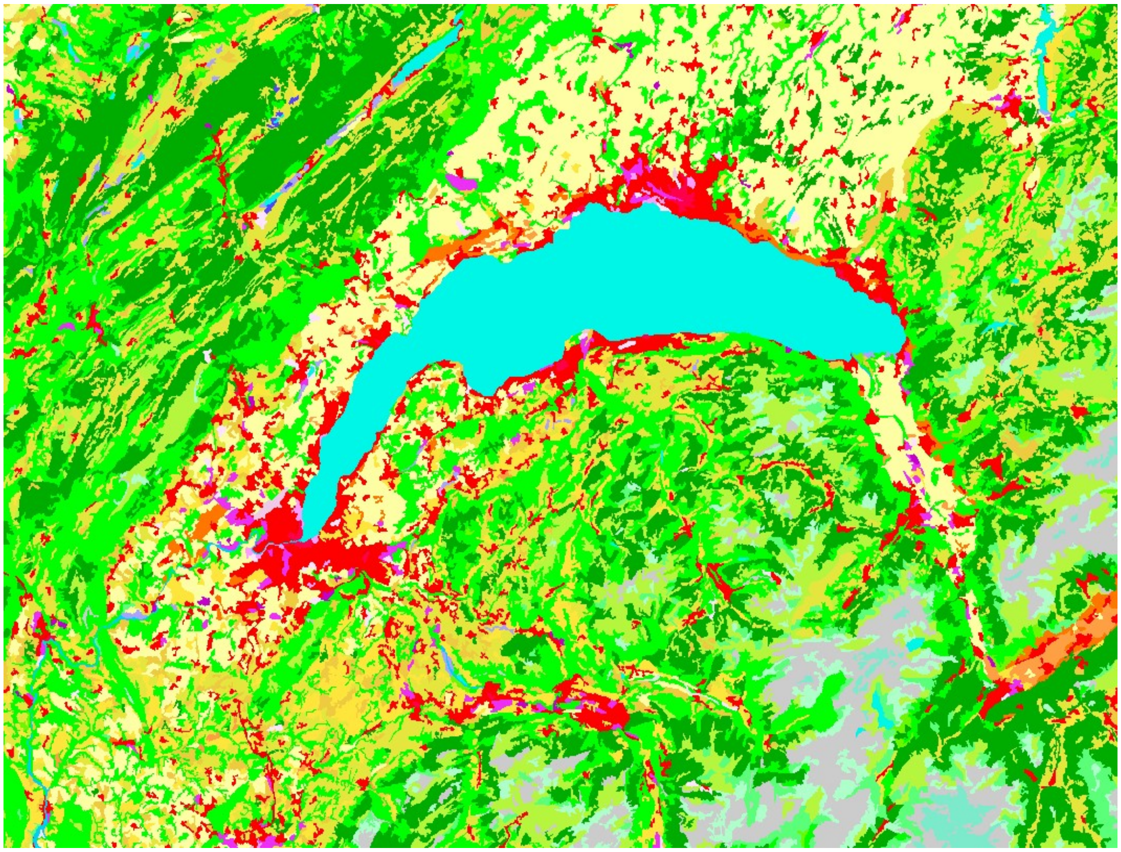


Figure 6-43: Corine data, an example of large-area landcover data derived from satellite images. These data include the area surrounding Lac Leman, the crescent-shaped feature in the center, and depict the relatively high resolution available, in this instance at a continental scale.

monitored and analyzed based on satellite data. Change detection often involves the combination of new images with previous landcover, infrastructure, or other information in spatial analyses to determine the extent of damage, to direct appropriate responses, and for long-range planning.

There are many examples of landcover data created across large areas, from states through continents. The Multi-Resolution Land Characteristics (MRLC) project aims to map landcover for the United States on approximately decadal frequencies, and the Corine project aims to map all of Europe at that or greater frequency (Figure 6-43).

Aerial or Satellite Images: Which to Use?

The value of satellite and aerial images for GIS should be clear. Several sources are often available for a given study area. An obvious question is "Which should I use?" A number of factors drive this choice.

First, the image data should provide the necessary spatial resolution. The resolving power of a system is generally defined by the smallest high-contrast object that can be detected, and is often approximately the pixel size. Current high-resolution satellite systems have effective spatial resolutions of 30 cm to several meters (foot to tens of feet). Images from digital mapping cameras, when taken at typical scales and with commonly used aerial scanners on planes, resolve objects in the 15 to 100 cm range (six inches to three feet). UAV-based imaging systems are often deployed to produce resolutions in the one to 10 centimeter range. Although the gaps are blurring, this high to lower resolution ladder still affects choice.

Second, the size of the analysis area should be considered. Aerial images are often less expensive for small areas. Aerial images are often available from government sources at low cost. Plane-based aerial images often cover from tens to hundreds of square kilometers, with low cost per square kilometer. As the size of the study area increases, the costs of using plane or UAV-acquired images may increase. Multi-image mosaics are often needed, raising costs, until at some point costs often surpass satellite images for the same area.

Third, satellite scanners may provide a broader spectral range and narrower bands relative to aerial images. As noted earlier, satellite scanners may detect well beyond the visible and near-infrared spectrums that are more common in aerial scanners. If important features are best detected using these portions of the spectrum, then satellite data are preferred. Broad-spectrum scanners are available for aerial and UAV systems, but these are rarer and tend to lose the cost advantage when compared to satellite systems. There are often tradeoffs between the size of the area to be imaged, resolution, and spectral bands used when selecting a system.

Finally, accuracy must be considered. Accuracy generally can't be much finer than the pixel resolution - you can't measure what you can't see. However, accuracies are often much poorer than image pixel sizes, and it is sometimes a grave mistake to assume image resolution and accuracies are equal. This is particularly true for UAV systems, where the GNSS systems used in drone positioning and for ground coordinate reference points may not be to as high a standard as with plane-based systems. Precise ortho-correction requires several advanced techniques, including and terrain and tilt removal, lens distortion removal, and ortho-registration. Professional-grade UAV systems are usually optimized for high accuracies. Less expensive, semi-professional or hobbyist UAV systems are common, and typically not designed for accuracy. While available at low cost, they often provide distorted data. A system should be chosen which provides both the resolution and accuracy required.

Airborne LiDAR

A number of laser-based, light detection and ranging systems (LiDAR) are becoming common. Lasers are pointed at the Earth's surface from an aerial or satellite platform, pulses of laser light emitted, and the reflected energy is recorded (Figure 6-44). Like radar, laser systems are active because they provide the energy that is sensed. Unlike radar, lasers have limited ability to penetrate clouds, smoke, or haze.

LiDAR systems have been used primarily to gather data about topography, vegetation, and water quality. Laser pulses reflect back from the canopy and the ground, and the strength and timing of the return is used to estimate ground height, canopy height, and other canopy characteristics (Figure 6-44). LiDAR signals over water also typically result in multiple returns, including water surface height and various depths, so lasers may be used to measure water clarity and nearshore water depth.

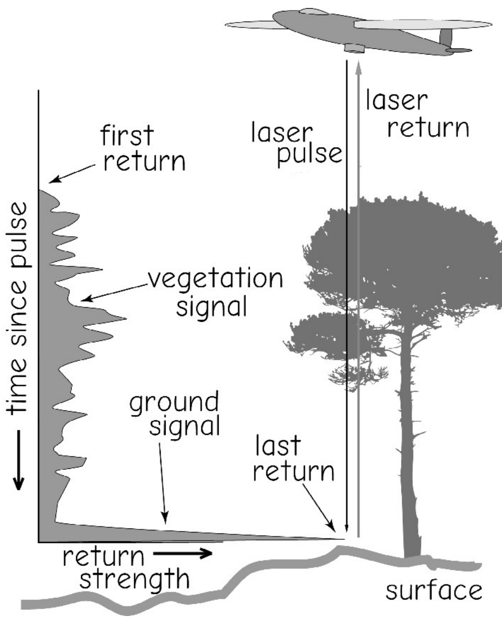


Figure 6-44: Laser mapping systems operate by generating and then sensing light pulses. The return strength is used to distinguish between vegetation and the ground, and the travel time may be used to determine heights.

Commercial LiDAR mapping systems are relatively new and have been used primarily for collecting surface data from aircraft and satellites. As noted earlier, three-dimensional LiDAR surveying from tripods or ground vehicles is growing, but we won't expand on them here.

Aerial LiDAR collection systems typically consist of a downward pointing LiDAR, a precision GNSS to record the plane's position to a very high accuracy, and an orientation sensing system to measure the angle of the LiDAR pulse relative to the vertical direction. LiDAR energy pulses are directed downward. Some energy from each pulse is reflected from vegetation, buildings, or other features above the ground, but under most conditions, many signals reach the ground and return to the airborne laser platform. The time interval between laser pulse emission and the ground return may be used to calculate aircraft height above the terrain. Flying height is known from the GNSS and the terrain elevation calculated for each pulse. Pulses may be sent several thousand times a second, so a trace of ground heights may be measured from every few centimeters to a few meters along the ground.

Discrete-return LiDAR is most common, wherein the system records specific values for each laser pulse downward. Typically, the first return from a pulse, last return, and perhaps one to several intermediate returns are recorded. *Waveform LiDAR* collects a continuous record of the pulse returns, the waveform trace shown in Figure 6-44.

Discrete-return LiDAR systems produce point clouds, consisting of X, Y, and Z coordinates (Figure 6-45), and the intensity, scan angle, return order, and other information. Modern laser systems often produce densities of several to tens of points per square meter of ground area, and these point clouds must be processed to remove errors, identify ground points, and assign points to feature types such as buildings or vegetation. Software for primary processing has been devel-

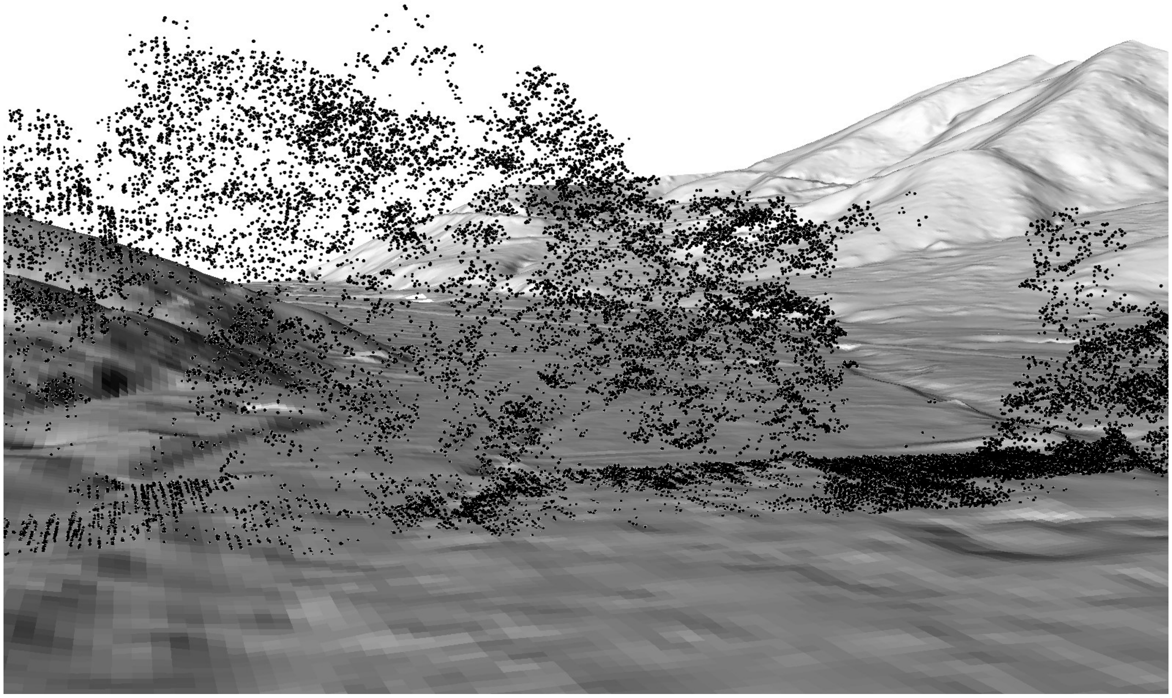


Figure 6-45: An example of a LiDAR point cloud, here a swath through a forested area, displayed over a terrain model. Each point represents a LiDAR return, showing returns from the canopy and taller trees, sub-canopy branches and shrubs, and dense ground returns in canopy gaps.

oped by most vendors so that files are delivered with the coordinate and height data assigned to the highest practical accuracy, and points classified with a standard number code that indicates the type of feature "hit by," or associated with each laser return. These standard codes identify ground (value = 2), buildings (value = 6), or water (value = 9). Several characteristics are used to classify points by feature type, including return strength, point order (first, last, or intermediate), local slope or texture, and the location and strength of adjacent returns.

There are a growing number of state-wide LiDAR projects, often driven by flood-plain mapping or for improved topographic measurements. Ground resolutions of 10 cm (4 in) or better are currently possible when LiDAR is combined with precise GNSS and aircraft orientation measurements. These projects report the "average" point density, but LiDAR returns are typically collected in swaths across the landscape, with individual scan lines discernible when viewed at large scales (Figure 6-46). Projects are planned and flown such that an appropriate amount of overlap exists between adjacent scans and adjacent flight paths, both to avoid gaps in coverage and areas with an unacceptably low sampling density.

Processing extracts the most relevant return for the desired product, for example, the maximum first return in a given square area may be extracted and assigned to a raster cell when calculating tree height, or a mean or minimum value when extracting ground heights. Different processing of the LiDAR point cloud will result in different extracted values.

Horizontal and vertical errors less than a few centimeters are attainable, allowing the use of airborne lasers to measure building height (Figure 6-47), floodplain location and extent, and slope and derived terrain characteristics, at much higher density and accuracy, over large areas, than previously possible.

LiDAR data have also been widely used to estimate vegetation characteristics, including tree height, forest density, forest wood amounts, growth rates, and forest type. A large number of points reach the ground in all but the densest forests, and the ground vs. locally highest canopy returns usually give an estimate of tree height that is as accurate as traditional manual measurements. The proportion of LiDAR returns is strongly related to canopy density, and to tree and forest wood mass. Crown shape can be determined from dense LiDAR data, which in turn helps separate forest types.

There is a standard LAS format, maintained by the American Society of Photogrammetry and Remote Sensing (ASPRS). The standard defines the file structure, content, storage order, naming, codes, and all other information so that any user may be able to access, process, and distribute LiDAR data in a standard way. The standard has evolved through various versions, up to 1.4 when this book edition was written. The convention defines the standard LiDAR exchange file with a .las file extension, for example, mylidar.las.

Also note that there are competing, non-standard, compressed formats defined by some providers, for example, ESRI supports their own "optimized" LAS format, and Rapidlasso has specified a different compressed format, with the .laz extension, used by the USGS National Map. Formats should at a minimum be openly defined, with all users having access to the file and storage specifications, and the ability to write independent code to read and write the files.

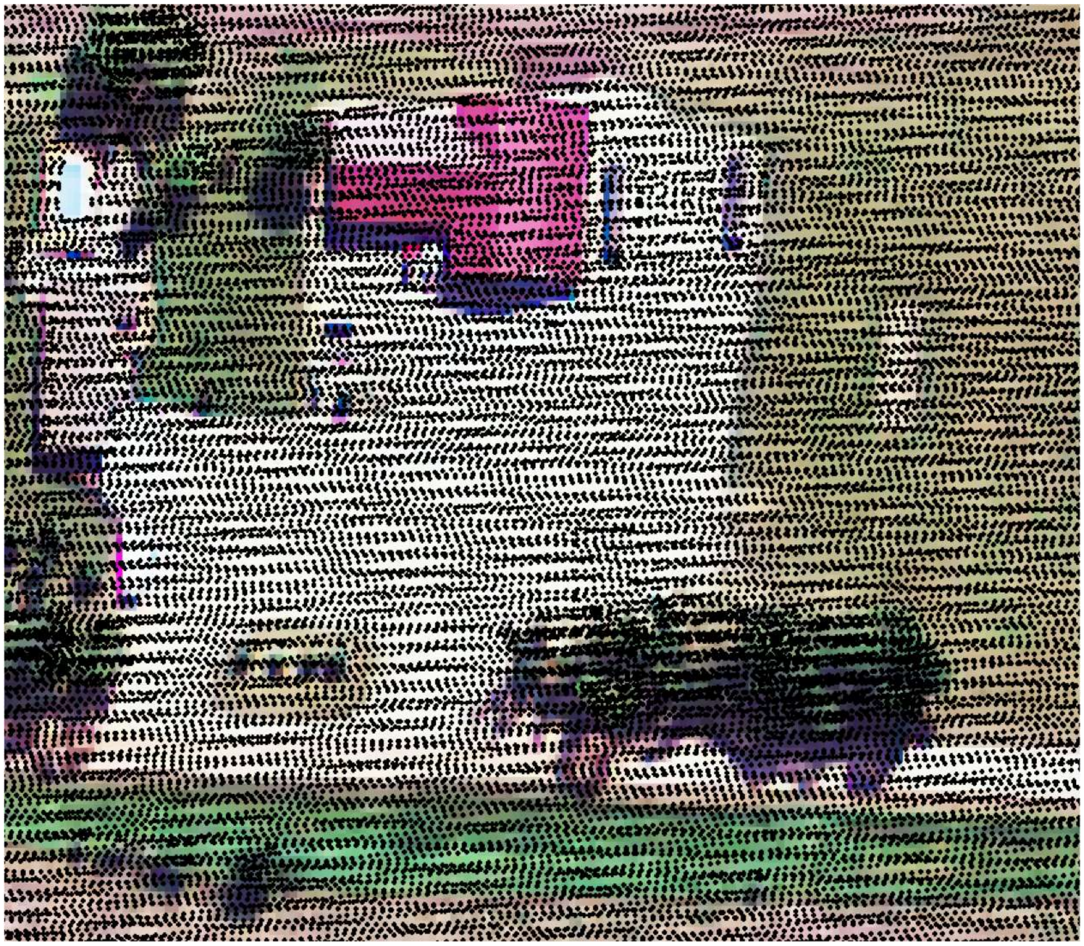


Figure 6-46: LiDAR sampling pattern, each dot represents a lidar return. Scan lines overlap, resulting in an uneven distribution of returns at very fine scales. Results are usually summarized for raster cells or areas that span gaps.

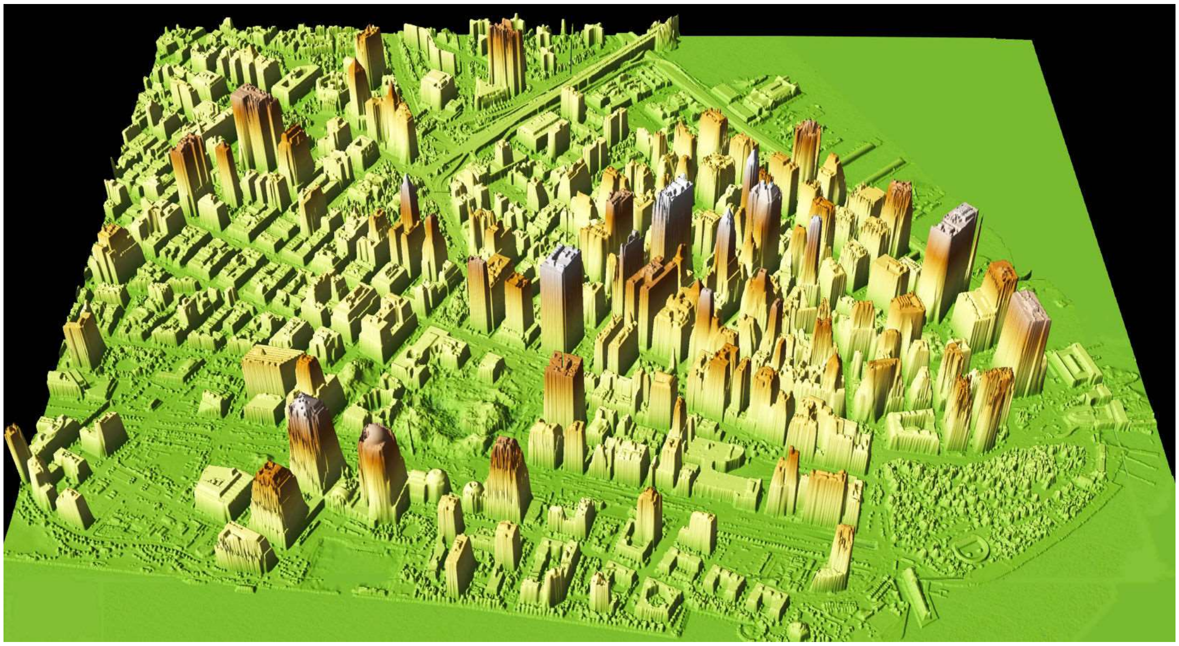


Figure 6-47: An example of LiDAR data and depiction of building heights. This image shows lower Manhattan, New York in late 2001. Tallest buildings are shown in white, and the land and water surfaces in green (courtesy NASA).

specifications, and the ability to write independent code to read and write the files.

Image Sources

National, state, provincial, or local governments are common sources of aerial images. These photographs are often provided at a reduced cost. For example, the National Agriculture Imagery Program (NAIP) provides coverage of much of the lower 48 United States on an annual basis. Images are usually collected in true color, but color infrared images may also be acquired, typically at a resolution of 1 meter or better. Photographs are usually collected during mid-growing season. The NAIP program is coordinated through the USDA Farm Services Administration, and so the images are sometimes referred to as FSA or FSA-NAIP photographs. Online and hardcopy indexes are available to aid in identifying appropriate image mosaics.

Aerial images may also be purchased from other government agencies or from private organizations. The U.S. Geological Survey (USGS) and U.S. Forest Service (USFS) routinely take aerial images for specialized purposes. The USFS uses aerial images to map forest type and condition, and often requires images at a higher spatial resolution and different time of year than those provided by NAIP. The USGS uses aerial images in the development of digital orthophotographs and maps. These organizations are also excellent sources of historical aerial images. Many government agencies contribute to a national archive of aerial images, some of which may be accessed via the internet.

Summary

Aerial and satellite images are valuable sources of spatial data. Photos and images provide large-area coverage, geometric accuracy, and a permanent record of spatial and attribute data, and techniques have been well developed for their use as a data source.

Remote sensing is based on differences among features in the amount of reflected electromagnetic energy. Chemical or electronic sensors record the amount of energy reflected from objects. Reflectance differences are the basis for images, which may in turn be interpreted to provide information on the type and location of important features.

Aerial images are a primary source of coordinate and attribute data. Camera-based mapping systems are well developed, and are the basis for most large-scale topographic maps currently in use. Camera tilt and terrain variation may cause large errors on aerial images; however, methods have been developed for the removal of these errors. Terrain-caused image displacement is the basis for stereo photographic determination of elevations.

Satellite images are available from a range of sources and for a number of specific purposes. Landsat, the first land remote sensing system, has been in operation for nearly 30 years, and has demonstrated the utility of satellite images. SPOT, AVHRR, Ikonos, and other satellite systems have been developed that provide a range of spatial, spectral, and temporal resolutions.

Aerial and satellite images often must be interpreted to provide useful spatial information. Aerial images are typically interpreted manually. An analyst identifies features based on their shape, size, texture, location, color, and brightness, and draws boundaries or locations, either on a hardcopy overlay, or on a scanned image. Satellite images are often interpreted using automated or semi-automated methods. Classification is a common interpretation technique that involves specifying spectral and perhaps spatial characteristics common to each feature type.

The choice of photographs or satellite imagery depends on the needs and budgets of the user. Aerial images often provide more detail, are less expensive, and are easily and inexpensively interpreted for small areas. Satellite images cover large areas in a uniform manner, and sense energy across a broader range of wavelengths.

LiDAR data are becoming a widespread source of spatial data. Discrete-return LiDAR are prevalent, providing X, Y, and Z coordinates for ground and above-ground feature returns. Most new, high-resolution digital elevation models are based on LiDAR data, and building and forest features are routinely extracted from LiDAR. Statewide acquisitions are becoming common, and system resolution and collection frequency are likely to improve through time.

Unmanned aerial vehicles (UAVs), also known as drones, show promise as spatial data collection tools. Lower costs, increased flexibility, and higher details must be weighed against limitations in throughput and hence area imaged, variability in accuracy, and regulatory uncertainty.

Suggested Reading

Atkinson, P., Tate, N. (1999). *Advances in Remote Sensing and GIS Analysis*. New York: Wiley.

Avery, T.E. (1973). *Interpretation of Aerial Photographs*. Minneapolis: Burgess.

Benoit, B., Ultre-Guerard, P. (2014). The NCES Earth observation program. *Geosciences and Remote Sensing Magazine*, 3:41–50.

Bolstad, P.V. (1992). Geometric errors in natural resource GIS data: the effects of tilt and terrain on aerial photographs. *Forest Science*, 38:367–380.

Brock, J.C., Purkis, S.J. (2009). The emerging role of Lidar remote sensing in coastal research and resource management. *Journal of Coastal Research*, 53:1–5.

Broderick, D.E., Frey, K.E., Rogan, J., Alexander, H.D., Zimov, N.S. (2015). Estimating upper soil horizon carbon stocks in a permafrost watershed of Northeast Siberia by integrating field measurements with Landsat-5 TM and WorldView-2 satellite data. *GIScience & Remote Sensing*, 52:131–157.

Campbell, J. B. (2006). *Introduction to Remote Sensing* (4th ed.). New York: Guilford.

Dial, G., Bowen, H., Gerlach, F., Grodecki, J., Oleszczuk, R. (2003). IKONOS satellite, imagery, and products. *Remote Sensing of Environment*, 14:23–36.

Ehlers, M. (1991). Multisensor image fusion techniques in remote sensing. *Journal of Photogrammetry and Remote Sensing*, 46:19–30.

Elachi, C. (1987). *Introduction to the Physics and Techniques of Remote Sensing*. New York: Wiley.

Fernandez-Diaz, J.C., Glennie, C.L., Carter, W.E., Shrestha, R.L., Sartori, M.P., Singhania, A., Legleiter, C.J., Overstreet, B.T. (2013). Early results of simultaneous terrain and shallow water bathymetry mapping using a single-wavelength airborne LiDAR sensor. *Selected Topics in Applied Earth Observation and Remote Sensing*, 7:623–635.

Fickas, K.C., Cohen, W.B., Yang, Z. (2015). Landsat-based monitoring of annual wetland change in the Willamette Valley of Oregon, USA from 1972 to 2012. *Wetlands Ecology and Management*, DOI 10.1007/s11273-015-9452-0.

Gillin, C.P., Cody, P., Bailey, S.W., McGuire, K.J., Prisley, S.P. (2015). Evaluation of Lidar-derived DEMs through terrain analysis and field comparison. *Photogrammetric Engineering and Remote Sensing*, 81:387–396.

Goetz, S.J., Wright, R.K., Smith, A.J., Zinecker, E., Schaub, E. (2003). IKONOS imagery for resource management: Tree cover, impervious surfaces, and riparian buffer analysis in the mid-Atlantic region. *Remote Sensing of Environment*, 8:195–208.

Hodgson, M.E., Bresnahan, P. (2004). Accuracy of airborne LiDAR-derived elevation: Empirical assessment and error budget. *Photogrammetric Engineering & Remote Sensing*, 70:331–339.

Lefsky, M.A., Cohen, W.B., Parker, G.G., Harding, D.J. (2002). LiDAR remote sensing for ecosystem studies. *Bioscience*, 52:19–30.

Lillesand, T.M., Kiefer, R.W., Chipman, J. (2007). *Remote Sensing and Image Interpretation* (6th ed.). New York: Wiley.

Mora, B., Tsendbazar, N.E., Herold, M., Arino, O. (2014). Global land cover mapping: Current status and future trends. *Remote Sensing and Digital Image Processing*, 18:11–30.

Nelson, R., Holben, B. (1986). Identifying deforestation in Brazil using multiresolution satellite data, *International Journal of Remote Sensing*, 1986, 7:429–448.

Olmanson, L.G., Brezonik, P.L., M.E. Bauer. (2014). Geospatial and temporal analysis of a 20-year record of Landsat-based water clarity in Minnesota’s 10,000 lakes. *Journal of the American Water Resources Association*, 50:748–761.

Pflugmacher, D., Cohen, W.B, Kennedy, R.E., Yang, Z. (2014). Using Landsat-derived disturbance and recovery history and lidar to map forest biomass dynamics. *Remote Sensing of Environment*, 151:124–137.

Richards, J.A., Jia, X. (2005). *Remote Sensing Digital Image Analysis.* New York: Springer.

Ryan, R., Baldbridge, B., Schowengerdt, R.A., Choi, T., Helder, D.L., Blonski, S. (2003). IKONOS spatial resolution and image interpretability characterization. *Remote Sensing of Environment*, 16:37–52.

Schowengerdt, R.A. (2006). *Remote Sensing: Models and Methods for Image Processing* (3rd ed.). New York: Academic Press.

Tigges, J., Lakes, T., Hostert, P. (2013). Urban vegetation classification: Benefits of multitemporal RapidEye satellite data. *Remote Sensing of Environment*, 136:66–75.

Warner, W. (1990). Accuracy and small-format surveys: The influence of scale and object definition on photo measurements. *ITC Journal*, 1:24–28.

Wolf, P.R., DeWit, B. (2000). *Elements of Photogrammetry with Applications of GIS*. New York: McGraw-Hill.

Woodcock, C.E., Strahler, A.H. (1987). The factor of scale in remote sensing. *Remote Sensing of Environment,* 21:311–332.

Yan, L., Roy, D.P. (2014). Automated crop field extraction from multi-temporal Web enabled Landsat data. *Remote Sensing of Environment*, 144:42–64.

Yang, C., Everitt, J.H., Bradford, J.M. (2006). Comparison of QuickBird satellite imagery and airborne imagery for mapping grain sorghum yield. *Precision Agriculture*, 7:33–44.

Study Problems and Questions

6.1 - Describe several positive attributes of images as data sources?

6.2 - What is the electromagnetic spectrum, and what are the principle wavelength regions?

6.3 - Define a spectral reflectance curve. Draw typical curves for vegetation and soil through the visible and infrared portions of the spectrum.

6.4 - Describe the structure and properties of digital sensors in digital aerial cameras.

6.5 - What are the basic components of a camera used for taking aerial photographs?

6.6 - Describe the most commonly used camera formats for aerial photography, and their relative advantages.

6.7 - What are the major sources of geometric distortion in aerial images, and why? What are other, usually minor, sources of geometric distortion in aerial images?

6.8 - What are typical magnitudes of geometric errors in uncorrected aerial images? How might these be reduced?

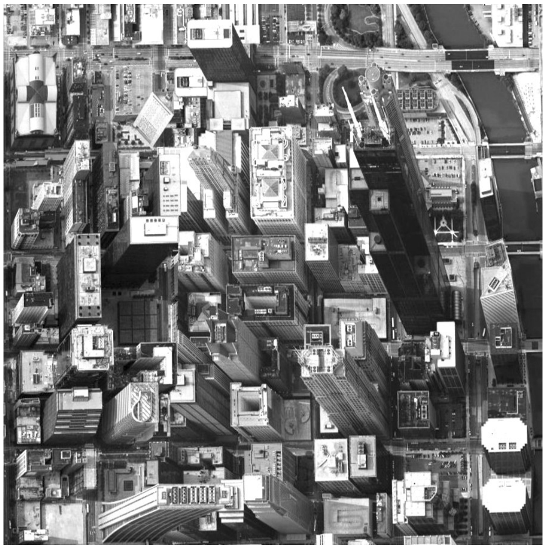
6.9 - A tall building is recorded on two vertical aerial photographs, the first photograph at a nominal scale of 1:20,000, the second photograph at a nominal scale of 1:40,000. The building is near the edge of both photographs, and terrain is level throughout the photograph. Which image will show a larger displacement, d, as shown in Figure 6-27?

6.10 - Describe stereo photographic coverage, and why it is useful.

6.11 - What is parallax, and why is it useful?

6.12 - Describe the basic process of terrain distortion removal.

6.13 - Why do the buildings lean in different directions in the images below?



6.14 - What is photointerpretation, and what are the main image characteristics used during interpretation?

6.15 - How are images from satellite scanners different from photographs? How are they similar?

6.16 - Describe and contrast the Landsat ETM+, SPOT, and WorldView-4 satellite imaging systems.

6.17 - What is a LiDAR? What type of information can LiDAR produce?

6.18 - What are three criteria used in selecting the type of images for spatial data development?

7 Digital Data

Introduction

Many spatial data currently exist in digital forms. Roads, political boundaries, water bodies, land cover, soils, elevation, and a host of other features have been mapped and converted to digital spatial data for much of the world. Because these data are often distributed at low or no cost, these existing digital data are often the easiest, quickest, and least expensive source for much spatial data (Figure 7-1).

Data are increasingly collected in digital formats. GNSS, laser measurements, and satellite scanners all provide primary data in digital forms. They are directly transferable to other digital devices and GIS systems, where they may be further processed. Direct digital collection should reduce transcription errors and help maintain source and processing history.

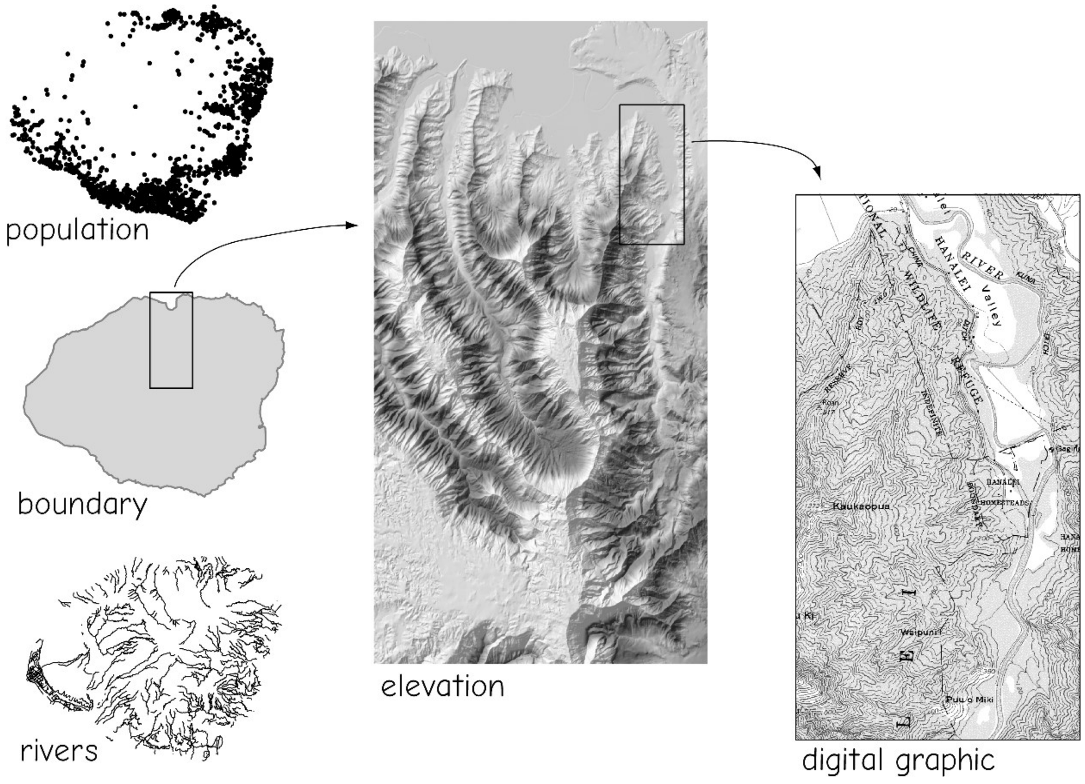


Figure 7-1: Examples of free digital data available at a range of themes, extents, and scales. Vector (left), raster (middle), and georeferenced digital graphic data (right) are shown for Kauai, Hawaii, USA.

Digital data are developed by governments because these data help provide basic public services such as safety, health, transportation, water, and energy. Spatial data aid disaster planning, national defense, and infrastructure development and maintenance. Many national, regional, and local governments have realized that once these data have been converted to digital formats for use within government, they may also be quite valuable for use outside government. Business, non-profit, education, and science, may draw benefit from the digital spatial data, as these organizations benefited in prior times from government-produced paper maps. Some data commonly available throughout the United States and the world are described in this chapter.

Map Services vs. Locally Stored Data

We must distinguish between data that are available for transfer to, storage on, and manipulation in a local computer (locally stored), from those data that are available as Web services, including Web Mapping Service (WMS), Web Feature Services (WFS), and Web Coverage Services (WCS). Digital data were first distributed on physical media, then via the Internet, but typically as electronic files that were copied onto a local storage device for use. You maintained a copy on your device, and manipulated those locally stored data. A WMS eliminates the need for a local copy.

A Web service is a standard way of serving geographic data over the Internet. GIS software access data via an Internet connection and display these data on a local machine, although they are "served" from some remote computing system. Image data are most often served, but vector data may also be provided, usually in the form of a georeferenced map backdrop. The data don't reside on the local hard disk, and data are delivered in response to each pan, zoom, or other change in display.

There are many differences among WMS, WFS, and WCS, but in broad strokes, a WMS is for serving cartographic data for producing and displaying maps, while WFS (vector) and WCS (primarily raster) deliver data and metadata in ways that ease spatial processing and analysis. Details of the differences are specified in standards documented by the Open Geospatial Consortium, www.opengeospatial.org/docs/is. Most data through services are currently provided as WMS, with few systems supporting and using the editing/analysis functions available through WFS and WCS.

Web services are better and worse than local data storage. Web services save space on the local hard drive, and only the portion of interest from a large data set need be accessed. A community of users may share the data, and the most up-to-date information provided to a wide set of users. Many different kinds of data may be joined together more easily, as accessing Web services typically requires a few mouse clicks. However, you may often not manipulate or change WMS data in any substantial way, and some kinds of analysis may not be supported or allowed. In these cases, local copies of the data may be required, or WFS or WCS developed. Map services may also require a fast and reliable internet connection, particularly for large raster data sets.

For the remainder of this chapter, we will primarily focus on data available for download, as these have the fewest barriers to use in analysis. Through time, many of these data may be offered via Web services, and software will ease use of Web-served data in analysis.

Global Digital Data

National governments commonly develop, organize, archive, and distribute national data sets. The standardization of weights and measures is a primary function of most national governments, and spatial data may be viewed as measurements of land, sea, or other national territories. Governments must oversee the planning, construction, and management of public infrastructure such as roads, waterways, and power distribution systems, and these activities, among many others, require spatial data sets that are national in extent.

National digital data are most often distributed via the Internet. For example, Geoscience Australia, a part of the Australian National Government, provides a suite of data layers that may be accessed via a website or may be requested in hardcopy form (Figure 7-2). Similar resources are available for Canada, most European countries, and many Asian and Latin American countries. A partial list of available data resources is included in appendix B, near the end of this book.

Global data sets are also available but are less common than national data sets.

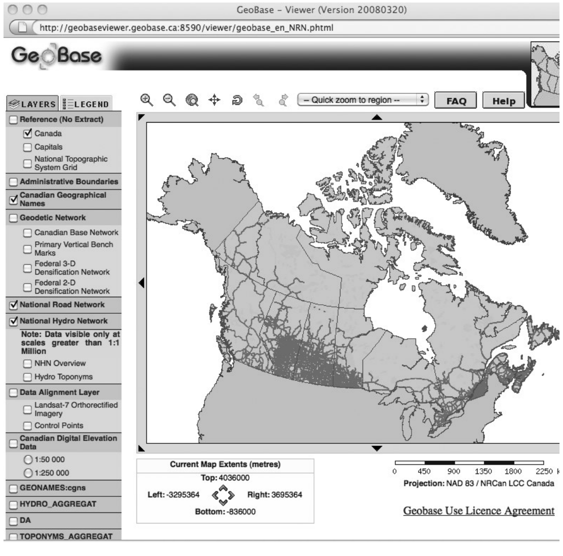


Figure 7-2: National governments often create portals through which digital data may be accessed.

Global data are scarce because few governments collect spatial data in the same way or with the same set of attributes. Different governments specify different datums, standard map projections, data variables, and attributes, or have different requirements for survey accuracy or measurement units. Data reduction or documentation methods may be different across national boundaries. There is substantial work in reconciling differences across national boundaries, therefore, global data sets are only occasionally built from a composite of national data sets.

A few global data sets are available that have been collected using a standard set of tools and methods. Global data sets have often been developed using global satellite data at a relatively coarse resolution (e.g., MODIS or VEGETATION canopy cover at one to eight kilometer cell sizes). Using a uniform global data source avoids the problem of reconciling differences among disparately collected data sets, but substantially reduces the number and type of global data sets that may be obtained. A limited set of data may be derived from satellite images. These features of interest must be visible from satellites, and there must be an organization interested in collecting and processing global data. NASA and the European Space Agency (ESA) provide a large and diverse group of global spatial data sets, due to their leadership in the development and application of satellite images at a range of spatial scales. Global raster data sets include elevation, land use, ecosystem type, and a number of measures of vegetation productivity, phenology, structure, and health.

University centers or ad hoc collaborations are other rich sources of global data. One example is the Center for International Earth Science Information Network, administered by the Earth Institute at Columbia University (www.ciesin.org). It seeks to provide global data to better address environmental problems. Another example is the Global Land Cover Facility at the Uni-

versity of Maryland (http://glcf.umd.edu), a set of Earth science data products, primarily derived from NASA satellites. A final example is the collection of Natural Earth data sets (http://www.naturalearthdata.com/), a volunteer collaboration for creating consistent, high-quality data suitable for small-scale mapping (Figure 7-3).

Global spatial data sets are often organized around a theme. For example, the Max Planck Institute in Germany has led an effort to create a gridded data set of historical global precipitation by combining data from 40,000 meteorological stations in 173 countries. These data are compiled, quality checked, and processed to create gridded data sets for normal precipitation. Data sets of annual anomalies, the number of gages, and systematic error are also provided. This was an expensive and time-consuming undertaking due to the number of different methods used to collect and report precipitation. Considerable time was spent reconciling data collection methods and results. A more complete description of these data is found at http://gpcc.dwd.de.

Global Spatial Data Infrastructure

Given the substantial difficulties in compiling data from disparate global sources, the Global Spatial Dataset Infrastructure (GSDI) initiative was formed. GSDI is an attempt to coordinate collection and processing methods worldwide to ensure that spatial data are broadly suitable for global-level analysis. The primary goal is to improve the development, use, and sharing of spatial data across the globe. This will be achieved through the adoption of common standards and complementary policies across governments and regions.

The GSDI initiative began in the late 1990s, and is still a work in progress. Activities during the first few years include identifying participants, developing goals and organizational structure, and identifying and prioritizing early actions. Activities on the GSDI initiatives may be found at www.gsdi.org.

The Global Map is one early GSDI initiative. The Global Map specifies common thematic layers: boundaries, elevation, land cover, vegetation, transportation, population centers, and drainage. Scale, feature classes, feature types, and feature names are

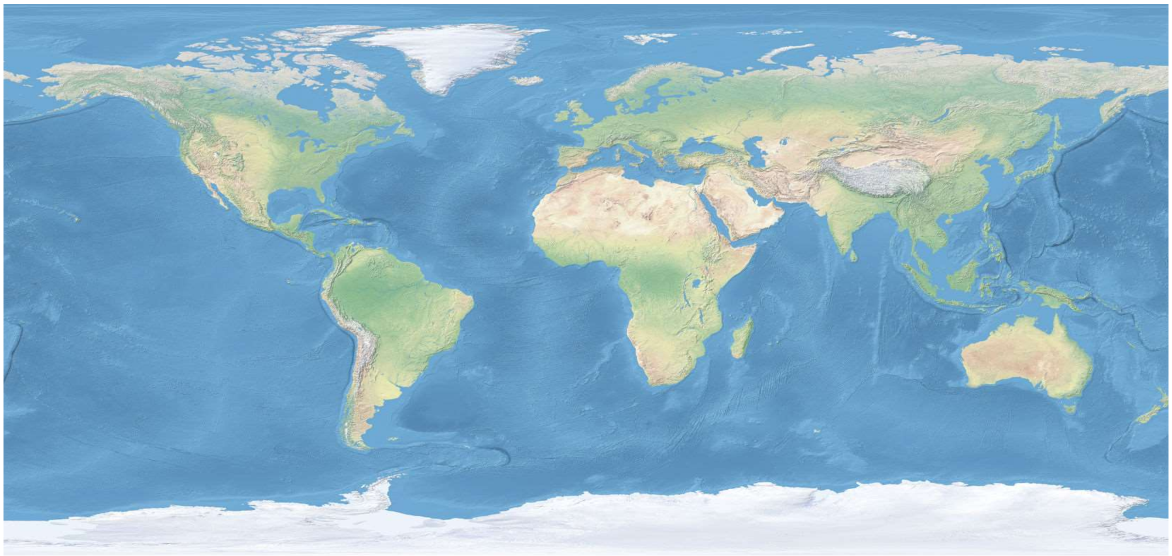


Figure 7-3: A diverse set of global spatial data types is available from various sources, although there is no one central clearinghouse. The United Nations, the European Union, and national government websites are the best sources, although there are specialized compilations by theme, as exemplified by the Natural Earth project, and available at http://www.shadedrelief.com/natural2/index.html.

Figure 7-4: An example of OpenStreetMap data for an area in Galicia, in northwestern Spain. Volunteer collaborators create line and attribute data representing important feature layers, at high detail.

specified, as are attributes, metadata, tiling schemes, and delivery mechanisms. Countries submit data to the Global Map project, which then serves as a distribution node.

Open Street Map

OpenStreetMap (OSM) is one notable effort to develop global data through international volunteer collaboration (Figure 7-4). Much like Wikipedia, this is an open access, user-generated resource. Individual users register and can check out data sets to modify. Roads and other transportation infrastructure are digitized, typically from image interpretation or via GNSS, and submitted for database integration. As with many online collaboratives, there are protocols for review and resolving conflicts, and data may be downloaded in various formats from OpenStreetMap or companion sites.These are often the best data in areas with poorly developed mapping infrastructure.

While OpenStreetMap provides the best data in many regions, there are potential drawbacks with these data. Because it is a collaborative, quality documentation and uniformity may be lacking. A range of sources, abilities, and methods may be used to develop data, and documentation on these sources may be unavailable. In addition, data may not be complete, depending on how much volunteer effort has been directed at an area, and the pace of change. Given these drawbacks, the data should be verified for accuracy and completeness, or at least suitability for the intended use, prior to adoption. This is true for all data, the burden perhaps falls more heavily on the user with crowd sourced data. Given the richness of detail of OSM data, it is well worth the effort.

Another perhaps slight barrier to use may be the method of distribution. Currently, the data may be downloaded from the primary website in a well-defined but little used data format. Data are available in more

standard formats from 3rd party services websites, and the native OpenStreetMap formats are supported by some softwares (e.g., QGIS), and will surely achieve broader support in the future. In spite of these potential drawbacks, these open-source collaborations have a bright future, and may well become standard for many types of data.

Other General Distributions

There are several global image archives, often focused on a specific satellite platform or initiative. For example, the Landsat system described in the previous chapter has been collecting data since the early 1970s. The LandLook initiative allows a global search for these data back to inception. LandLook also supports Sentinel-2 data search, browse, and download. Similar archives exist for SPOT and other long-running, government funded platforms.

NASA hosts diverse set of data through their Land Products Data Active Archive Center (DAAC). They include various global digital elevation data sets, among them the Shuttle Radar Topography Mission (SRTM) and the ASTER satellite global elevation data. There is a comprehensive archive of layers derived from MODIS satellites, including global vegetation density and type, forest cover change, phenologies, and various physical measures such as, albedo and surface reflectances.

ESRI Open Data is another rich source of global data, containing a broad range of categories. In early 2019, over 115,000 different data sets were hosted, to view and download, depending on permissions. Data are available for political boundaries, demography, education, health, agriculture, economic variables, natural resources, and other categories. Metadata, links to downloads and programmer's display interfaces, and webmap production are provided. While the collection is U.S. weighted, there are many international data sets.

The NASA-funded Socioeconomic Data and Applications Center (SEDAC) provides substantial data with a global focus. Population, urban land cover and characteristics, global agricultural lands and food supply, roads, environmental resources use and sustainability are among the data sets served. Most data involve a combination of remote sensing and other data collection efforts in global or regional estimation, hence the support by NASA. Data are free to download in various standard formats, with metadata and development methods defined.

The United Nations Environment Program (UNEP) distributes over 500 global data sets through the Environmental Data Explorer. Data are searchable, provided at national, UNEP regional, or subregional units of area, and downloadable in standard formats. Data focus on environmental themes, broadly defined.

Terra Populus is another global spatial data portal, focusing on integrated population and environmental data. It provides matching data downscaled to compatible areas, including land use, land cover, and climate data.

There are many other general distribution sites serving global data, both general and specific. These are best discovered in the subject matter literature, and via broad web searches.

Digital Data for the United States

National Spatial Data Infrastructure

The United States has defined the National Spatial Data Infrastructure (NSDI) as the policies, technologies, and personnel required to ensure the efficient sharing and use of spatial data. The goal of the NSDI is to reduce duplication of effort among agencies, to improve quality and reduce the costs of geographic information, to make geographic data more accessible to the public, to increase the benefits of available data, and to establish key partnerships with states, counties, cities, tribal nations, academia, and the private sector to increase data availability (www.fgdc.gov).

The NSDI has developed a framework that identifies core data sets commonly used by many organizations. The framework consists of geodetic control, orthoimagery, elevation, transportation, hydrography, cadastral data (property boundaries), and governmental unit boundaries. A primary goal of the NSDI is to foster the efficient development of these core data.

The NSDI advocated parallel access to many data sets across a range of government agencies. Geoplatform.gov is one result of U.S. federal government efforts, providing shared geospatial data, web services, and applications. The U.S. Geological Survey (USGS), http://www.usgs.gov/, is another good source of geospatial data from the U.S. federal government, and many of these data will be described in the following sections.

The U.S. National Map

Digital data are available for most of the United States, through the National Map project, described as a cornerstone of U.S. mapping efforts (http://nationalmap.gov). Data are provided on political and civil boundaries, transportation, hydrography, geographic names, structures (e.g., dams, notable buildings, towers, or monuments), elevation, aerial photographs, and land cover (Figure 7-5). Some of these data are available from other dedicated projects and websites, for example, elevation datasets (USGS 3DEP program and website) and the National Hydrologic Datasets (USGS NHD program and website). We'll discuss these two data sources in detail in later sections of this chapter, and here focus on a general description of the National Map project and the additional data available through the national map.

The National Map also distributes transportation, structures, and boundaries data in vector formats. Transportation data represent roads, railroads, airports, and other transportation features. National map boundaries data identify national, state, county, and Native American lands, as well as the boundaries for cities and towns. Structures data identify selected man-made facilities, including government centers or service buildings, hospitals, and other important buildings, as well as dams, bridges, and other key physical structures. Limited attribute data are provided with all the National Map vector data.

Data for the National Map come from a variety of sources, including new primary data collections from aerial and satellite images contributed by federal and state agencies, and older data. Much of the National Map data are legacies of USGS hardcopy mapping programs. The USGS began topographic mapping in the United States in the 1880s, but the most common detailed map series began in the 1940s, with the production of 1:24,000 scale topographic maps. These paper maps covered 7.5 minutes of arc on a side, and comprised about 55,000 tiles covering the lower 48 states. These data were converted to digital layers, known as Digital Line Graphs (DLGs), and many DLG data are still available on legacy websites, although the updated National Map data should be used where available.

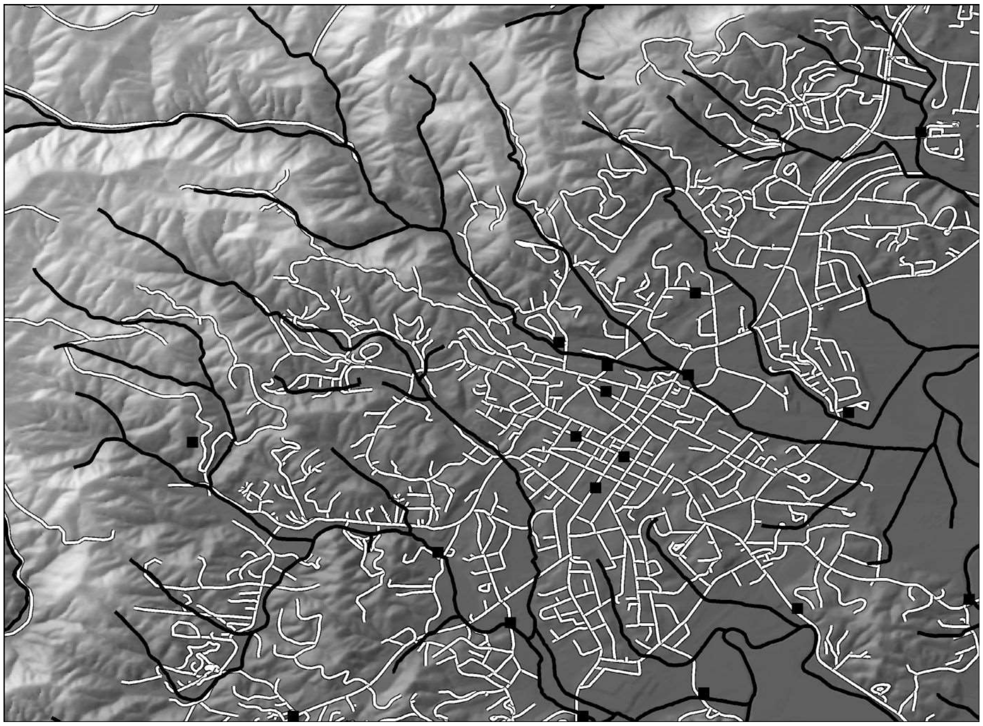


Figure 7-5: An example of spatial data available through the U.S. National Map, here elevation, road, river, and government building data for an area near Brevard, North Carolina.

The mapping program aimed at paper quadrangle maps ended in the 1990s, and has been replaced by a digital format USTopo map (Figure 7-6). Currently, these digital topographic maps are delivered as geographically enhanced postscript document format (PDF) files, with layers for orthoimagery, roads, place names, elevation contours, and rivers, lakes, and other hydrographic features. Layers may be rendered visible or invisible, and the maps displayed with other georeferenced data in appropriate viewers, but these maps are generally not used in a GIS. Complete USTopo maps are not available for all of the 7.5 minute cells for the lower 48 states, but are planned.

Figure 7-6: An example of a USGS USTopo map.

Digital Elevation Models

Digital elevation models (DEMs) provide elevation data in a raster format and are available at 10 m or better resolution for all states but Alaska, and are used commonly in data analysis and display (Figure 7-7). DEM manipulations and terrain analysis are described in detail in Chapter 11, but for now, know they are among the most used spatial data sets in many endeavors. Here we introduce the sources of U.S. DEM data and their basic characteristics.

Ground and aerial surveys are the primary source of original elevation measurements for most DEMs. Traditional distance and angle measurements with surveying equipment were used up until the 1940s to provide precise elevations at specified locations. Because these methods are relatively slow, they provided a sparse network of points, with a dense network suitable for elevation mapping over only small areas. Improved electronic distance meters helped, as did global positioning system technologies, but even with these improvements in survey speed and accuracy, these technologies are too slow to be the sole elevation data collection method over all but the smallest areas.

Aerial images and airborne LiDAR surveys complement field surveying by increasing the number and density of measured elevations. Accurate elevation data may be collected over broad areas with the appropri-

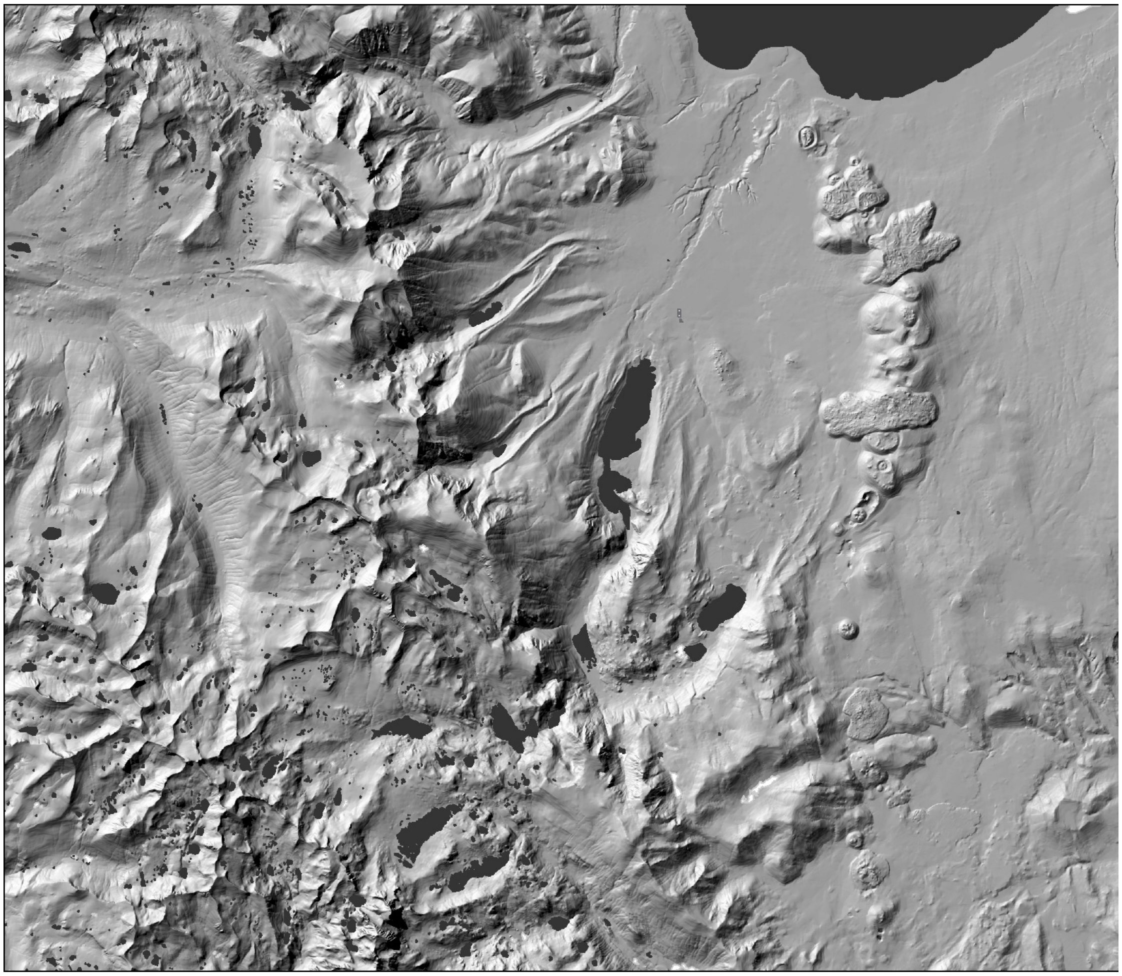


Figure 7-7: Digital elevation models (DEMs) are available at various resolutions and coverage areas for most of the world. This figure near Mono Lake in eastern California was produced using data available from the USGS. A string of lava flows is apparent on the right side of the image, and glacial lakes and valleys to the left.

ate selection of aerial mapping technologies. From the 1950s until the late 1990s, most elevation data were compiled using precise mapping aerial cameras, complemented by optical ground surveys. Since the late 1990s, LiDAR mapping has been combined with GNSS to more accurately and rapidly map elevation. These various GNSS and survey methods are discussed in Chapter 5, and aerial images and LiDAR in Chapter 6. Laser-based elevation mapping and DEM generation are now common, and will be used for the foreseeable future to create the highest-resolution DEMs.

In areas where LiDAR DEMs are not yet available in the United States, most DEMs have been developed using photogrammetry, precision measurements from metric aerial photographs. Photogrammetry has been used since the 1930's to map lines of a constant elevation (contours) and spot heights, and data have been developed for much of the Earth's surface using these techniques.

DEMs with 3-, 10-, and 30-meter horizontal sampling frequency are available for much of the United States. Currently, the USGS delivers these DEMs as part of the 3D Elevation Program (3DEP), through the USGS National Map portal. Data are available at a 30-meter resolution for almost all of the United States, and 10-meter resolution for the lower 48 states and Hawaii, and at 1 and 3 m for a large and expanding area. The underlying LiDAR data are also available for download. Note that global 30 m SRTM and Aster data, described in the global data section of this chapter, are available for the United States, but are generally inferior to the 10 m or better resolution data where both exist. There is a special 5 m radar-based data set restricted to Alaska.

GIS users should be cautious because there are several versions of DEM data for many areas, and they should generally use the more current, higher accuracy, or higher-resolution data. The existence of various elevaton data sets, covering various time periods, does provide the opportunity to monitor change through time, for example, broad-scale mining modifications in Kentucky (Figure 7-8). The most current USGS data are best accessed through the USGS Earth Explorer or National Map portals.

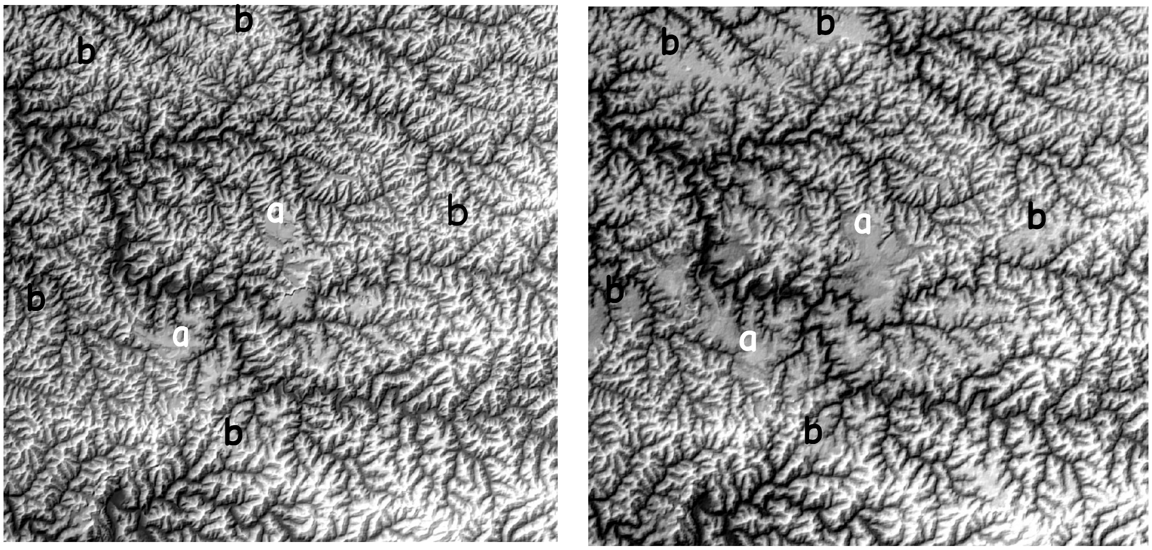


Figure 7-8: The extent of mountaintop removal strip mining in eastern Kentucky is evident in this comparison of older NED data (left) and year 2000 SRTM elevation data (right). Ridgetops are light colored, valley bottoms dark colored. Mines appear as areas of uniform tone on ridgetops. Mine sites labeled a, above, have expanded substantially, while large new mines have also been developed (b).

Hydrologic Data

The National Hydrologic Dataset (NHD, and NHD Plus) contains digital spatial data about surface waters, including rivers, streams, canals, ditches, lakes, ponds, springs, and wells. The NHD combines data from USGS digital line graph data and United States Environmental Protection Agency (EPA) river *reach* data. A reach is a segment of a stream, river, or coastline considered homogenous under an EPA classification scheme. NHD data are based on 1:100,000 scale USGS DLG data, but may be improved as new data are developed. Naturally occurring and built features are represented in NHD data (Figure 7-9). These include water bodies, canals, pipelines, dams, and other natural or control structures. Attributes may be provided for these features, for example, a lake type or name, if a dam is earthen or concrete, or ditch type. Features may be points, lines, or polygons.

NHD data also represent network topology, the connection among stream features, and include information on connections and flow directions. Line segments have a designated flow direction, and connections or crossings may be represented as full connections, or noted as a bypass, for example, when a spillway or pipeline crosses a river without the possibility of discharging water into the river. Coding schemes have been developed to identify each reach in the hydrographic network, and to represent network connections among reaches.

NHD data are organized by areas, in a hierarchically nested set of *Hydrologic Units*, identified by unique codes (HUCs). These units correspond to watersheds, or basins, or logical aggregations or subareas of watersheds (Figure 7-10). The United States was divided into 21 regions, and these regions further divided into 222 subregions. Subregions were in turn divided, forming a total of 352 hydrologic units, and these are further divided into 2,150 hydrologic units. This fourth level division is for the most part along major river basins, outlining distinct watersheds, or intermediate pieces along a the main stem of larger rivers. Each of these divisions is identified by a unique eight digit code, and so these areas are also known as

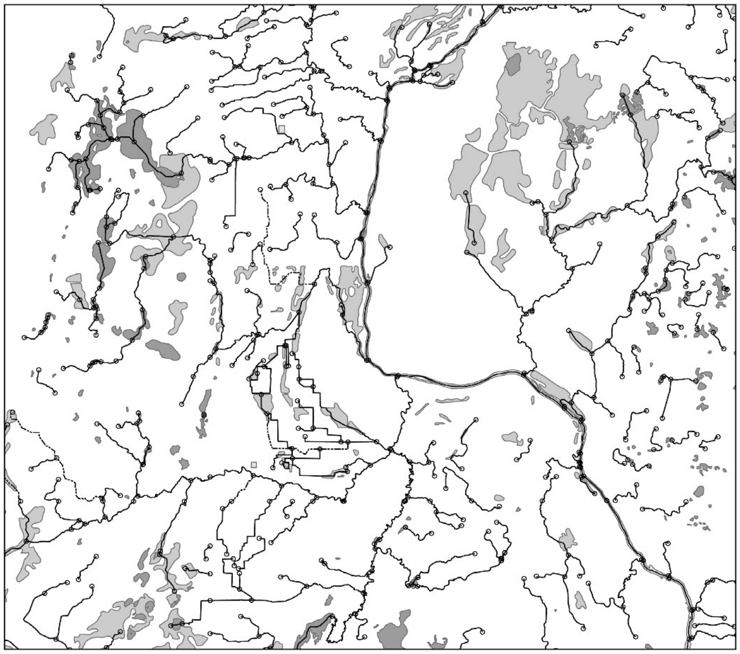


Figure 7-9: An example of sub-basin data obtained from the National Hydrologic Dataset. A number of feature types are represented, including stream segment endpoints (unfilled circles), connected stream networks (solid lines), water bodies (dark polygons), and adjacent wetlands (grey polygons).

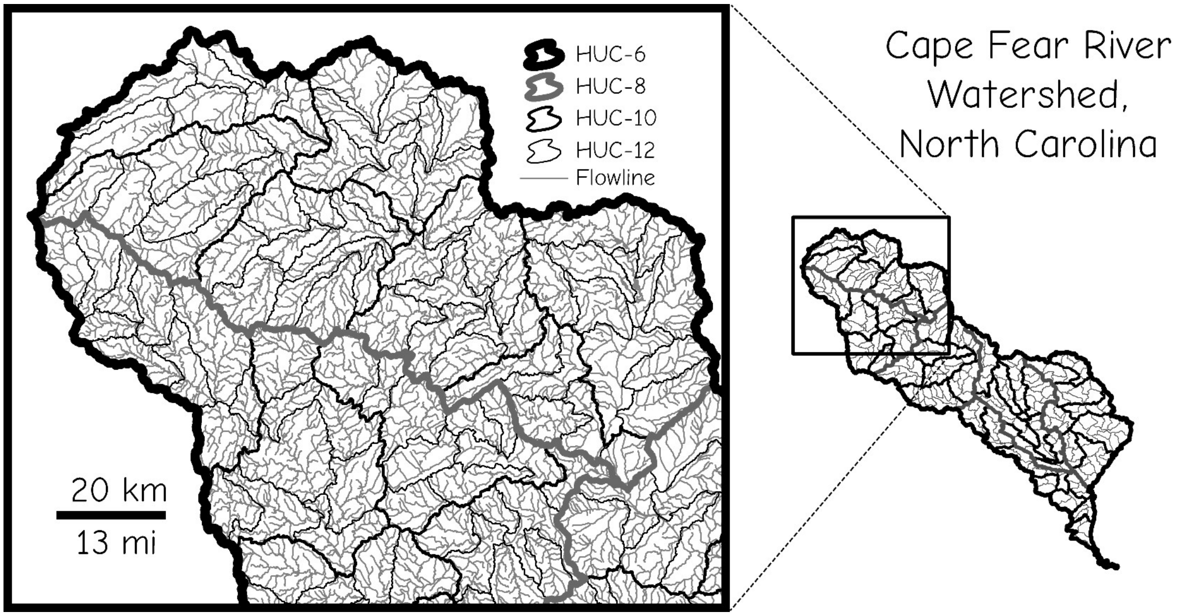


Figure 7-10: An example of nested HUC drainage areas for a portion of the Cape Fear River in North Carolina. Higher HUC designators nest within lower designators, so HUC-12 areas are contained within HUC-8 units.

HUC-8 boundaries. Regions, subregions, and subregion divisions are known as HUC-2, HUC-4, and HUC-6 areas, respectively, providing a nested set of drainage areas for areas larger than the HUC-8 catchments. HUC-8 catchments may in turn be split into smaller HUC-10 and HUC-12 catchments, this last size typically the smallest delineations widely available.

The United States EPA also provides data on waters and watersheds of various types and formats, organized to correspond to the HUC data at some levels. EPA River Reach Files organize data in a series of versions, from RF1 through RF3 data. RF3 data are designed to provide a nationally consistent hydrographic database that records geography and assigns unique identifiers to all surface water features. It allows the hydrologic ordering of reaches so that larger rivers and segments may be accurately defined, along with river connectedness and flow direction. RF3 data also record the locations and characteristics of additional elements, including gages, dams, and other hydrologic features.

River reach data are precursors to NHD data, and so contain much of the same base information. RF files are available for most of the contiguous United States. Tabular data on water chemistry and other watershed characteristics are available at http://cfpub.epa.gov/surf/locate/index.cfm.

There are other improved hydrologic data, called NHDPlus and NHDPlus HR. Managers need consistent elevation, stream, and watershed boundary data sets at high resolution to solve many water resource science and planning problems. The original NHD produced in the early 2000s focused on hydrographic data from 1:100,000 scale. Subsequent work has focused on improving the accuracy, consistency, and tools to support NHDPlus data, and with subsequent versions using improved digital elevation data and enforcing consistency with other data sources.

There is an emerging system for storing, finding, and retrieving hydrologic data associated with the CUAHSI project (www.cuahsi.org). CUAHSI is a National Science Foundation funded project involving more than 120 universities to support hydrologic science and education. CUAHSI-HIS is an internet-based system for sharing hydrologic data via a Web service. As noted

earlier, a Web service is a set of protocols that allow communication among computer programs over the internet. In GIS these Web services are most often used to streamline data sharing and access. The CUASHI HIS is designed to aid the integration and sharing of disparate hydrologic data, such as stream gauge, precipitation, river location, basin topography or other basin characteristics.

There are other hydrologic data sets available for the United States, more generalized, and most often used for analysis or display over larger areas. The USGS produced digital, nationwide data sets based on paper 1:100,000, 1:250,000, 1:1,000,000, and smaller-scale maps (Figure 7-11), and these are available from various state, national, local, and private sources. These data show larger rivers and a limited set of attributes for each river, most importantly river names. These data are also not hydrologically continuous, in that many of the rivers do not maintain their connection through water bodies. Despite these limitations, they are often used because they may be more appropriate for statewide or regional analysis involving only the main stems or larger rivers in a region.

Integrated, consistent, continent to worldwide hydrography data are also available from the "Natural Earth" data projects (www.naturalearthdata.com). These data are intended for use in cartography, and not primarily for analysis, as they have been generalized and made consistent primarily for display rather than geographic accuracy. Different hydrographic data are offered, targeted at a range of small scales, for regional through global mapping. A limited set of attributes is available, including reach or river names and cartographic widths.

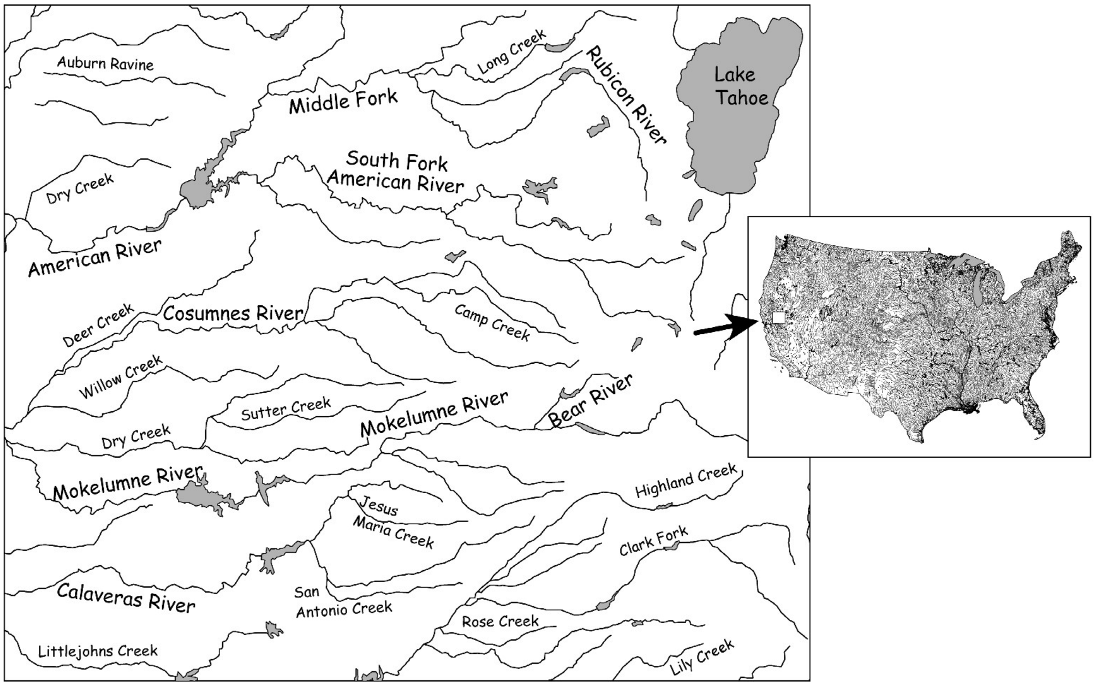


Figure 7-11: An example of USGS legacy 1:1,000,000 hydrologic data, with the national data set (inset, right) and a portion in east-central California depicting the level of detail in these data.

High-Resolution Digital Images

Digital images are available from a range of sources, including national, state, and county governments, or from private contractors, satellite imaging companies, and resellers. High-resolution digital image data are typically collected every five to ten years by the USGS, in partnerships with states or other government agencies. Image archives go back to the 1950s for most of the United States, with state and countywide coverage as far back as the 1930s. Nation-wide coverage was completed in the 1980s and 1990s through the NHAP and then NAPP programs, at scales in the 1:40,000 to 1:60,000 range. These formed a primary basis for digital ortho photo quads, or DOQQs, the first orthographic, high-resolution digital images with national coverage. Most other images from before the early 2000s were film-based, although some have been scanned and are available from the United States EROS Data Center. Image sets include the historical black and white aerial photographs, nation-wide programs of the 1980s and 1990s, high-resolution coastal images, radar, and other special collections.

The High-Resolution Orthophotographs (HROs) series are among the highest resolution, widely available image data sets. The HRO images are collected and distributed through the USGS with current coverage for about 1/3 of the U.S. These data are often collected at 0.3 m (1 ft) resolution, and at times up to 10 cm (4 in) resolution. Because they are orthophotographs, object base locations have been corrected for tilt and terrain distortion at ground height. Towers, buildings, bridges, and other tall objects often appear to tilt, as these structure heights above ground are not corrected on the images (Figure 7-12). These images are use-

Figure 7-12 An example of a High-Resolution Orthophotograph (HRO), distributed by the U.S. Geological Survey. Individual persons can be identified near the flagpole in the upper left part of the island, as well as the statue and shadow to the lower right.

ful for infrastructure mapping, planning, disaster management, and many other applications. The images are sometimes used for vegetation mapping, but these images are most often collected during leaf-off periods, and so must often be complemented by images taken during leaf-on periods for most vegetation mapping efforts.

The HRO and other high-resolution images are valuable sources of spatial data. These images are typically processed to within a few pixels of the delivered resolution, for example, typically accurate to within a half-meter for the 0.3 m data. Because these and other photos described below record the surface at a fixed point in time, they may be used to create new maps or to monitor change (Figure 7-13).

NAIP Digital Images

The National Aerial Imagery Program (NAIP) acquires photographs during the growing season in the continental United States. NAIP images are distinct from the previous HRO, NHAP, NAP, and DOQ programs because NAIP is primarily for one purpose — to monitor agricultural landscapes. NAIP photographs are typically acquired during the full-leaf period for local crops, so the bulk of the images are collected from June through August, in contrast to other photographic programs, which were often taken during leaf-off conditions. In addition, the NAIP photographs typically have a yearly repeat cycle, while other sources are often spaced at five-year or longer intervals. NAIP photographs may be obtained in hardcopy or digital formats, commonly as county mosaics. Data may be viewed from within a GIS using a publicly accessible web service, where data are stored centrally, rather than on a local computer disk, as described earlier in this chapter. These data may then be used as a backdrop for digitizing, wherein the analyst extracts information through a visual classification of spatial features.

Figure 7-13: An example of a historical aerial photograph from the 1940s (left) and 2008 (right), for an area in east-central Minnesota. Early development was restricted to near the lake along the top margin of the 1940 photo, while by 2008 the area had become completely suburbanized. Photographs may be used to map current infrastructure and resources, and their change through time.

Images are most often collected as natural color, digital aerial photographs, although sometimes infrared bands are also collected. NAIP images are orthorectified and provided at 1- and 2-meter ground resolutions, with corresponding horizontal accuracies at 5 to 10 m. Data are typically provided in an NAD83 UTM coordinate system corresponding to the image area.

NAIP images are most useful as a base for digitizing, particularly when information on vegetation type or condition is important (Figure 7-14). Leaf-on NAIP images are more useful for mapping vegetation, because differences are most often expressed in the color, brightness, and texture of foliage. While the natural color images typically used for NAIP images are inferior to infrared images, substantial information on vegetation can be collected, and sometimes the NAIP images include an infrared band. This, plus the annual image collection cycle, make these images a valuable source of spatial data.

National Land Cover Data

While land cover is important when managing many spatially distributed resources, data on land cover are quite expensive to obtain over large areas. These data are often scarce, at low categorical or spatial resolution, and rarely available over broad areas. While individual states, counties, metropolitan areas, or private landholders have developed detailed land cover maps, there have been few national efforts to map land cover in a consistent manner. There are four consistent national data sets available, based on satellite data from a consistent set of categories, and a legacy data set from the 1970s and early 1980s.

The National Land Cover Database (NLCD) is the most recent and detailed source of national land cover information. NLCD versions are produced in a cooperative effort by a number of United States federal government agencies, under the Multi-Resolution Land Characteristics Consor-

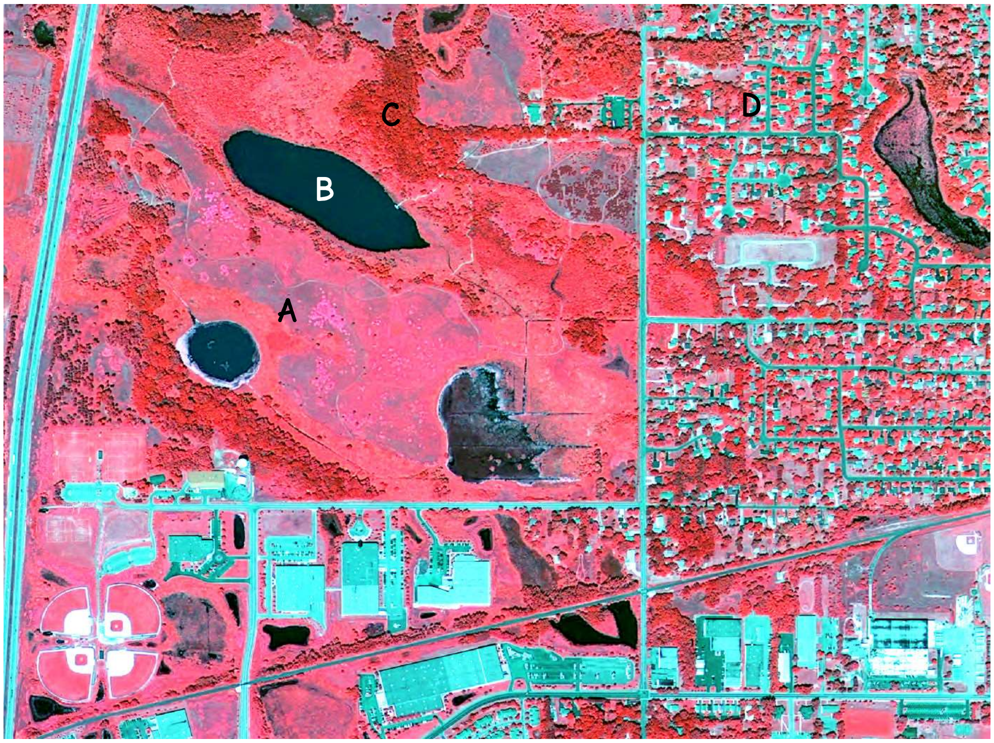


Figure 7-14: An example NAIP image showing wetlands (A), lakes (B), forest (C), and residential areas (D). These images are particularly useful for land cover and land use assessments and detecting fine-detail changes at annual time steps.

tium (MRLC), so these are sometimes referred to as MRLC data. The consortium's goal is a consistent, current land cover data record for the conterminous United States. NLCD data have been produced four times, known respectively as NLCD 1992, NLCD 2001, NLCD 2006, and NLCD 2011, corresponding to the primary data collection years (Figure 7-15).

NLCD land cover classifications are based primarily on 30 m Landsat Thematic Mapper data. NLCD 1992 land cover was assigned to one of 21 classes. Full coverage is obtained from adjacent or overlapping, cloud-free Landsat images. Multiple dates are often acquired in order to improve accuracy and categorical detail through phenologically driven changes. For example, evergreen forests are more easily distinguished from deciduous forests when both leaf-off and leaf-on images are used. Other spatial data sets are used to improve the accuracy and categorical detail possible through spectral data alone. These data include digital elevation, slope, aspect, Bureau of Census population and housing density data, USGS LULC data, National Wetlands Inventory data, and STATSGO soils data.

Data are processed in a uniform manner within each state or region, and a national set of categories and protocols is followed. All classifications were subjected to a standardized accuracy assessment, and reported and delivered in a standard format. Accuracy assessments were based on NAPP or other medium- to high-resolution aerial photographs or, in later versions, with high-resolution satellite data. Areas were stratified based on the images, and sampling units defined. Photointerpretations of land cover were assumed true, and compared to NLCD classification assignments. Errors were noted and reported using standard methods.

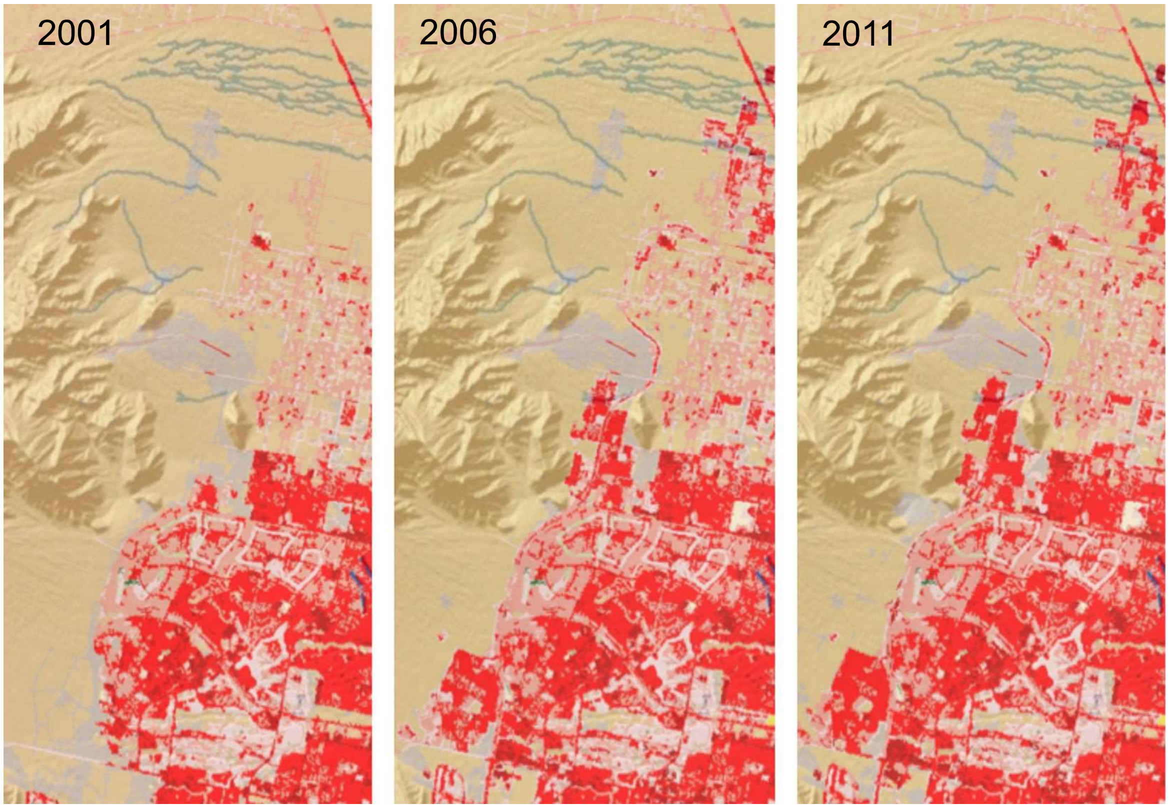


Figure 7-15: A time series of National Land Cover Data (NLCD) for an area north of Las Vegas, NV. NLCD categorizes land cover into 21 classes, and are provided in a 30-meter raster cell format. In this increasing urbanization is shown in the spreading tones of darker red, with agricultural areas in lighter tones of gray (courtesy MRLC).

NLCD 2001 analysis was refined to yield more categories, higher accuracy, and a more uniform classification. Landsat data from three periods, digital elevation data, population density, road locations, NLCD 1992, and city lights data were used (Table 7-1), and previous data recoded to create a consistent time series. The base data were also used to estimate percent impervious surface, and tree canopy density. NLCD 2006 and 2011 again refined methods, incorporating information on previous classifications, maintaining categories but improving accuracy and uniformity.

A 2016 NLCD/MRLC data set is under development. It promises to update the classification, and was not out in early 2019, at the time of this writing.

NASS CDL

The National Agricultural Statistical Service (NASS) produces yearly Crop Data Layer (CDL) data, land cover maps that focus on distinguishing major crop types and rotations (Figure 7-16). These data are created from a combination of existing land cover data for nonagricultural lands, multidate images from mid-resolution satellites such as Landsat and Resourcesat-1, coarse resolution but higher frequency MODIS data for phenological discrimination, and various vector data layers to improve classification accuracy.

CDL land cover classification is based on extensive field surveys conducted by the United States Department of Agriculture. Fields are visited, airphotos obtained, and fields, farms, and regions classified by dominant crop types and rotations. Observed crop types are compared to spectral data from sat-

Table 7-1: NLCD 2011 land cover classes. Classes have varied slightly through versions.

Water	***Shrubland***
11 open water	51 dwarf shrub
12 perennial ice/snow	52 shrub/scrub
Developed	***Herbaceous Upland Natural***
21 developed, open space	71 grassland/herbaceous
22 developed, low intensity	72 sedge/herbaceous
23 developed, medium intensity	73 lichens
24 developed, high intensity	74 moss
Barren	***Herbaceous Planted/Cultivated***
31 bare rock/sand/clay	81 pasture/hay
	82 cultivated crop
Forested Upland	***Wetlands***
41 deciduous forest	90 woody wetlands
42 evergreen forests	95 emergent herbaceous wetlands
43 mixed forests	

ellites, and a classification algorithm developed. Classification methods have changed since 2002, the year nationwide data became available annually. Class assignment accuracies are generally between 85 and 95% for agricultural crops.

CDL data is produced annually for most regions, allowing analysis of trends in planting, crop rotations, and harvest. Data may be downloaded for regional, statewide, or sub-state areas, in standard formats and coordinate systems

While NASS-CDL data are the most up-to-date and accurate land cover classification for agricultural lands, they have limitations. Classification for nonagricultural lands are not as rigorously ground truthed as agricultural data, and depend on older NLCD classifications. Land cover is classified only for counties with agriculture, although this is a surprisingly large proportion of the country. The 30 to 60 m cell size is quite good for such a large-area classification, but still too small for field-level assessments, and is better suited to farm-level and larger analyses.

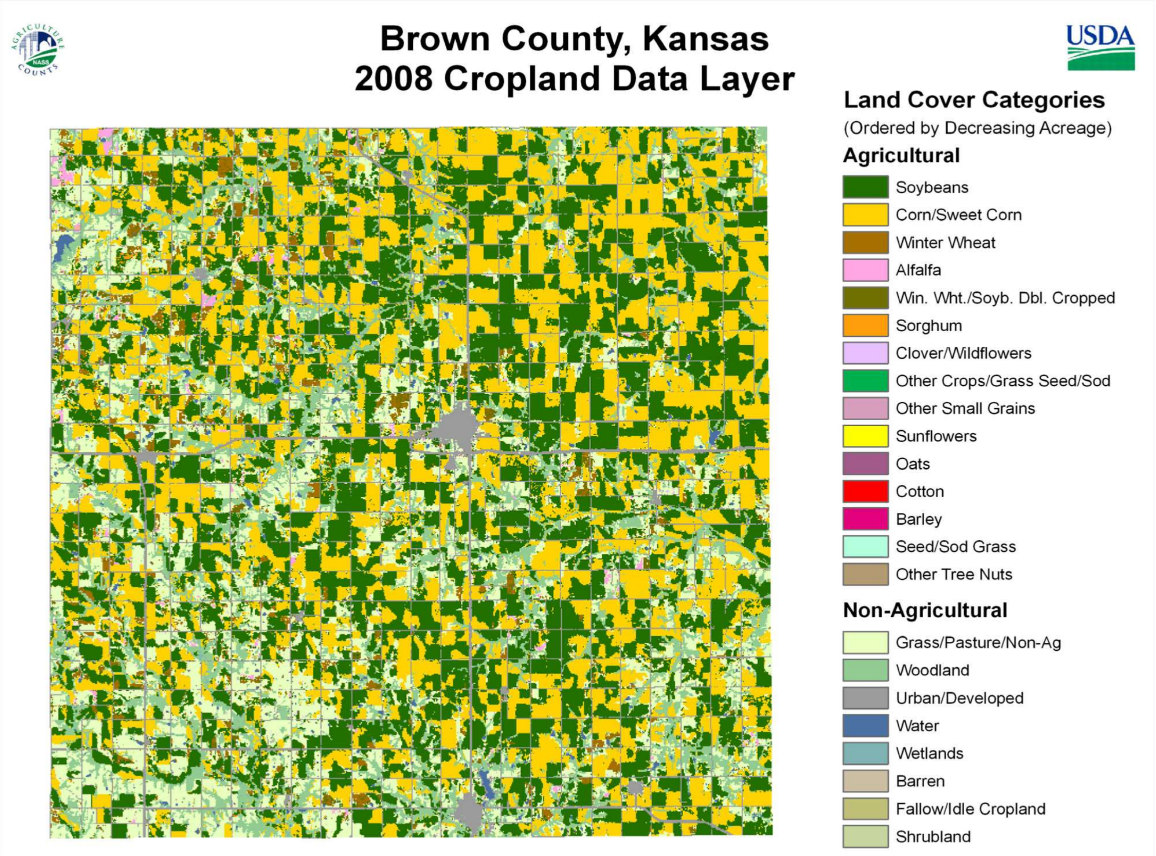


Figure 7-16: An example of NASS Crop Data Layer information on agricultural land cover, here for a region in eastern Kansas. The pattern is dominated by row crops, chiefly soybeans and corn, the rectangular blocks that increase in size toward the northeastern portion of the figure (courtesy USDA).

National Wetlands Inventory

Data on the location and condition of wetlands are available for much of the United States through the National Wetlands Inventory (NWI) program. NWI data are produced by the United States Fish and Wildlife Service. NWI data portray the extent and characteristics of wetlands, including open water (Figure 7-17), and are available for approximately 90% of the conterminous United States. About 60% of the conterminous United States is available in digital formats. NWI data were produced through the 1970s and 1980s, with an update in the 1990s. Decadal updates are planned.

NWI data were produced through a combination of field visits and airphoto interpretation. Spring photographs at a range of scales and types are used. Color infrared photographs at a scale of 1:40,000 were commonly used; however, black and white photographs and scales ranging between 1:20,000 and 1:62,500 have been employed. Spring photographs typically record times of highest water tables and are most likely to record ephemeral wetlands. There is substantial year-to-year variability in surface water levels, and hence there may be substantial wetland omission when photographs are acquired during a dry year.

NWI data provide information on wetland type through a hierarchical classification scheme, with modifiers. Wetlands are categorized as part of a lacustrine (lake), palustrine (pond), or riverine system. Subsystem designators then specify further attributes, to record if the wetland is perennial, intermittent, littoral, or deep water. Further class and subclass designators and modifiers provide additional information on wetland characteristics. A shorthand designator is often used to specify the wetland class. A wetland may be designated L1UB2G, as system = lacustrine (L),

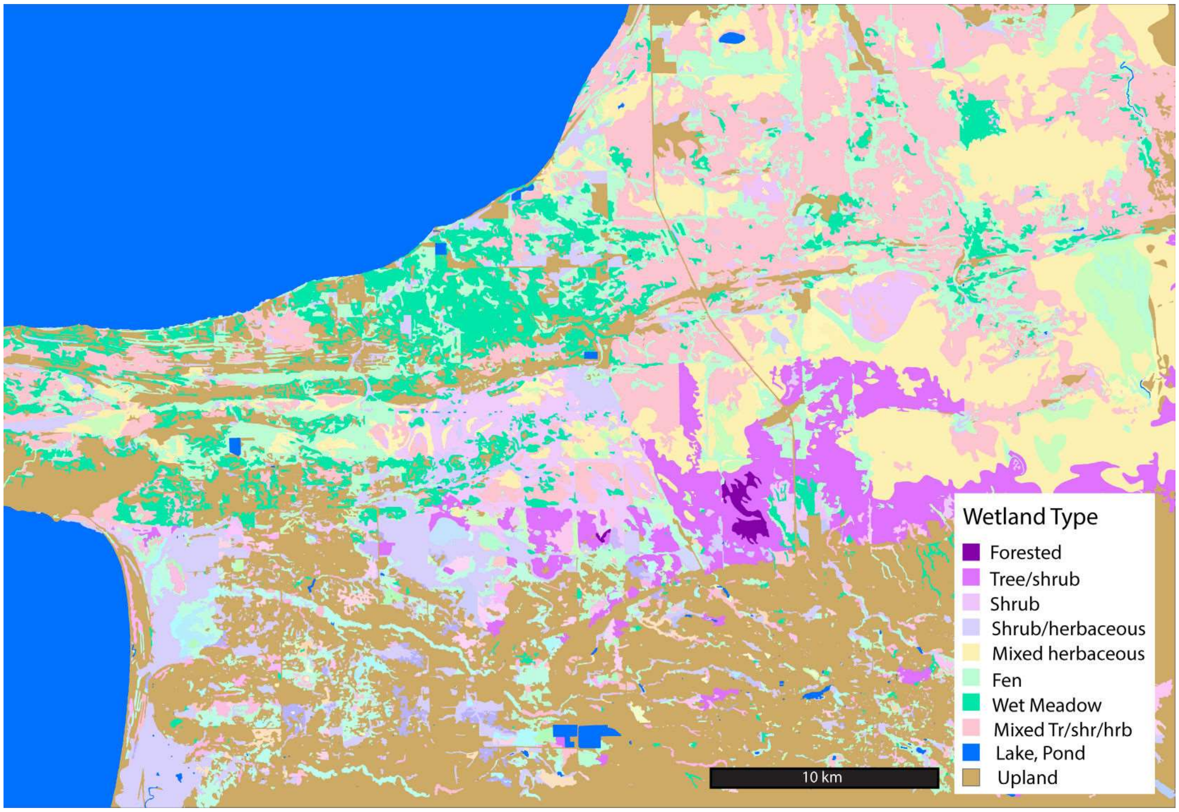


Figure 7-17: An example of national wetlands inventory (NWI) data. Digital NWI data are available for most of the United States, and provide information on the location and characteristics of wetlands.

subsystem = limnetic (1), class = unconsolidated bottom (UB), with subclass = sand (2), and a modifier indicating the wetland is intermittently exposed (G).

The *minimum mapping unit* (MMU) is the target size of the smallest feature captured. Features smaller than the MMU are not recorded in these data. NWI data typically specify MMUs of between 0.5 and 2 ha. MMUs vary by vegetation type, film source, region, and time period. MMUs are typically largest in forested areas and smallest in agricultural or developed areas, because it is more difficult to detect many forested wetlands. MMUs also tend to be larger on smaller-scale photographs. The MMU, scale, and other characteristics of the wetlands data are available in map-specific metadata.

NWI data do not exhaustively define the location of wetlands in an area. Because of the photo scales and methods used, many wetlands are not included. Statutory wetland definition typically includes not only surface water, but also characteristic vegetation or evidence on the surface or in the soils that indicates a period of saturation. Since this saturation may be transient or the evidence may not be visible on aerial photographs, many wetlands may be omitted from the NWI. Nonetheless, NWI data are an effective tool for identifying the location and extent of large wetlands, the type of wetland, and for directing further, more detailed ground surveys.

Digital Soils Data

The Natural Resource Conservation Service (NRCS) of the United States Department of Agriculture has developed three digital soils data sets. These data sets differ in the scale of the source maps or data, and thus in the spatial detail and extent of coverage. The National Soil Geography (NATSGO) data set is a highly generalized soils map for the continental United States, developed from small-scale maps. NATSGO data have limited use for most regional or more detailed analyses and will not be further discussed here. State Soil Geographic (STATSGO) data are intermediate in scale and resolution, and Soil Survey Geographic (SSURGO) data provide the most spatial and categorical detail.

SSURGO data are intended for use by land owners, farmers, and planners at the large farm to county level. SSURGO maps indicate the geographic location and extent of the soil map units within the soil survey area (Figure 7-18). Soil map units typically correspond to general grouping, called phases, of detailed soil mapping types. These detailed mapping types are called soil series. There are approximately 18,000 soil series in the United States, and several phases for most series, so there are potentially a large number of map units. Only a small subset of series is likely to occur in a mapped area, typically fewer than a few hundred soil series or series phases. A few to thousands of distinct polygons may occur.

SSURGO data are developed from a combination of field and photo-based measurements. Trained soil surveyors conduct a series of field transects in an area to determine relationships among soil mapping units and terrain, vegetation, and land use. Aerial photographs at scales of 1:12,000 to 1:40,000 are used in the field to aid in location and navigation through the landscape. Soil map unit boundaries are then interpreted onto aerial photographs or corresponding orthophotographs or maps. Typical photo scales are 1:15,840, 1:20,000, or 1:24,000. These maps are then digitized in a manner that does not appreciably affect positional accuracy. Soil surveys are often conducted on a county basis, so county mosaics of SSURGO data are common. SSURGO data are reported to have positional accuracy no worse than 13 m (43 ft) for approximately 90% of the well-defined points when SSURGO data are compiled at 1:24,000 scale.

SSURGO data are linked to a Map Unit Interpretations Record (MUIR) attribute database (Figure 7-19). Key fields are provided with the SSURGO data, including a unique identifier most often related to a soil

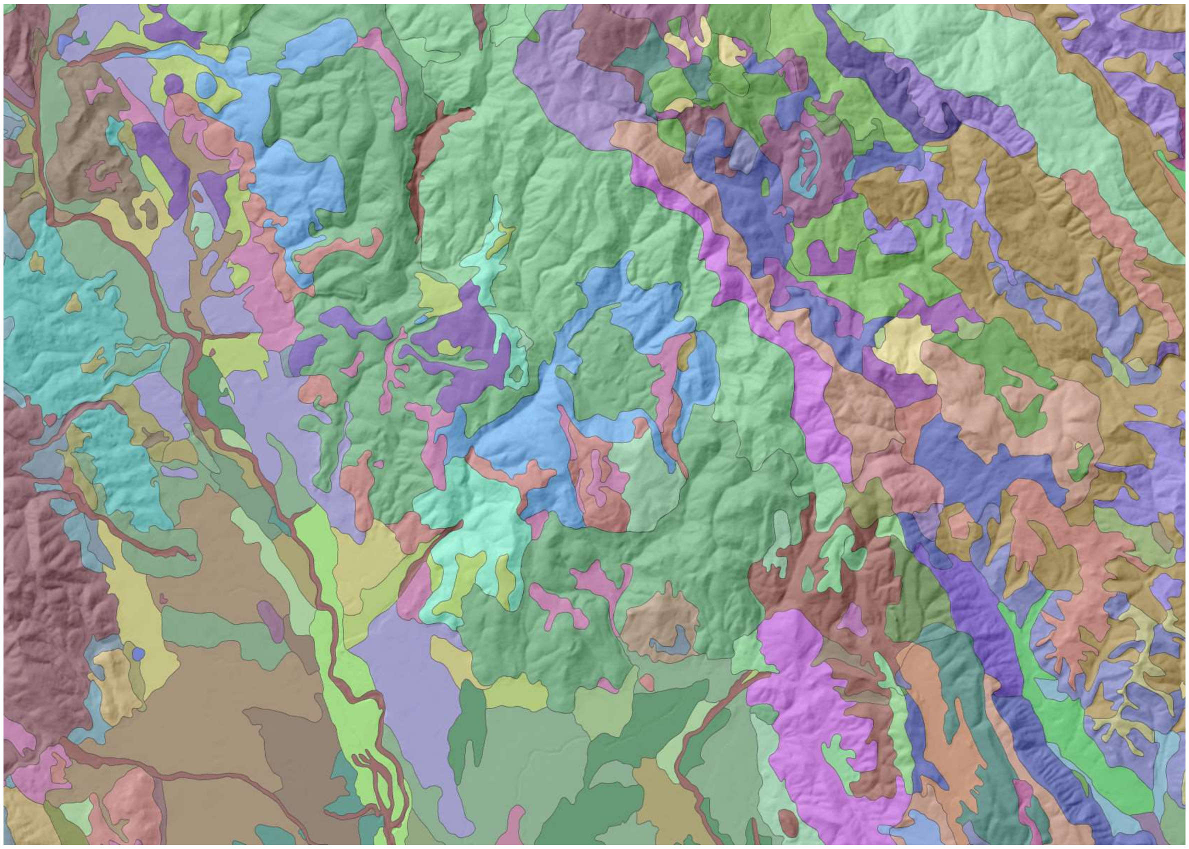


Figure 7-18: An example of SSURGO digital soils data available from the NRCS. Each polygon represents a soil mapping unit of relatively uniform soil properties.

map unit, known as the map unit identifier (muid). Tables in the MUIR database are linked via the muid, and other key fields. Most tables contain the muid field, so a link may be created between the muid value for a polygon and the muid value in another table, such as the Compyld table (Figure 7-19). This creates an expanded table that may be further linked through cropname, classcode, or other key fields. These kinds of table structures and linkages are discussed in Chapter 8.

Variables include an extensive set of soil physical and chemical properties. Data are reported for water capacity, soil pH, salinity, depth to bedrock, building suitability, and most appropriate crops or other uses. Most MUIR data report a range of values for each soil property. Ranges are determined from representative field-collected samples for each map unit, or from data collected from similar map units. Samples are analyzed using standardized chemical and physical methods.

STATSGO digital soil maps are also available, at a smaller scale and over broader areas than SSURGO soil data. STATSGO data are typically created by generalizing SSURGO data. STASTGO map units are larger, more generalized, and do not necessarily follow the same boundaries as SSURGO map units. In addition, STATSGO polygons contain from one to over 20 different SSURGO detailed map units. Each STATSGO map unit may be made up of thousands of these more detailed SSURGO polygons, and many different SSURGO map unit types can be represented within a STATSGO polygon. STATSGO data provide information on some of this variability. Data and properties on multiple components are preserved for each STATSGO map unit.

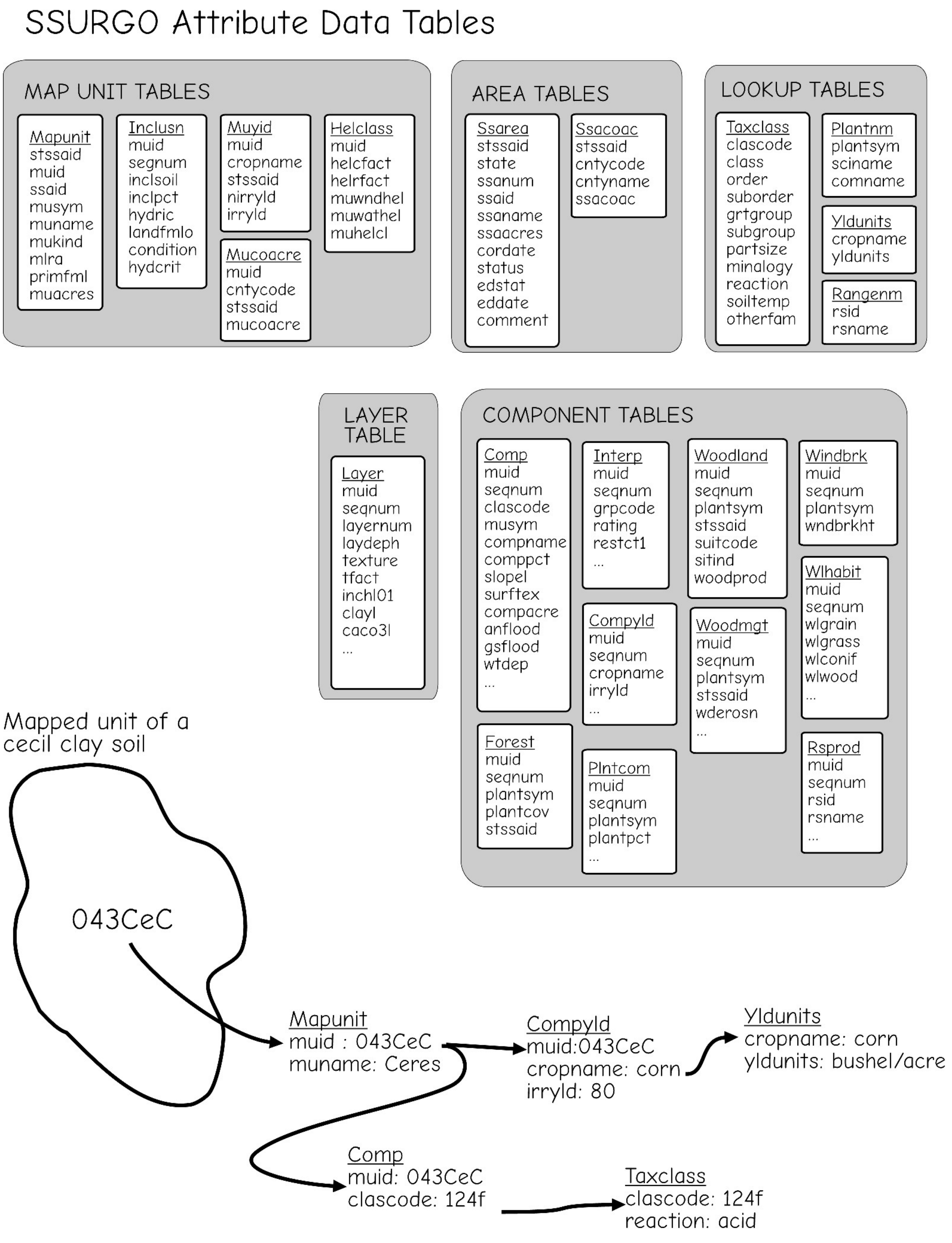


Figure 7-19: The database schema associated with the SSURGO digital soils data. Variables describing soil characteristics are provided in a set of relatable tables. Keys in each table, shown in bold, provide access to items of interest. Codes provided with the digital geographic data, e.g., the muid, provide a link to these data tables. The relation of a mapped soil polygon to attribute data is shown in the example at the bottom. The muid is related from the MAP UNIT and COMP tables, which in turn are used to access other variables through additional keys.

Digital Floodplain Data

Floods cause billions of dollars in damage each year in the United States; losses could be reduced with the effective application of GIS. A first step is the mapping of flood-prone areas. The Federal Emergency Management Agency (FEMA) develops and disseminates flood hazard maps, commonly known as floodplain maps (Figure 7-20). These maps locate the boundary of areas with a 1% or higher annual chance of flooding, commonly known as 100-year floodplain maps.

FEMA occasionally updates these floodplain maps. The objectives are to develop maps of flood hazard via an improved process, with better input data, in a uniform digital format, and to integrate map creation into ongoing local and state government mapping and planning efforts. Updates are particularly important given the changes in precipitation intensity and with sea level rise due to changing climates.

Floodplain maps are used for a number of purposes, chief among them setting flood insurance rates. Over 19,000 communities participate in the National Flood Insurance Program (NFIP). This supports federal government to guarantee flood insurance for communities with floodplain management ordinances. Ordinances reduce flooding risks for redevelopment and new construction, thereby reducing losses.

Digital floodplain maps are produced to define regions within a 100-year floodplain. Boundary accuracy may be challenged, usually by individual landowners, businesses, or municipalities, and the proposed adjustments evaluated and included in the floodplain map if they improve accuracy.

Maps are most often produced by cooperating technical partners (CTP). Expertise, training, and demonstrated capabilities are required of each CTP. Best technical practices for digital floodplain data development are defined by FEMA, and training is offered to teach best available methods and increase data quality. Protocols for verifying and revising maps are defined as part of the data evaluation process.

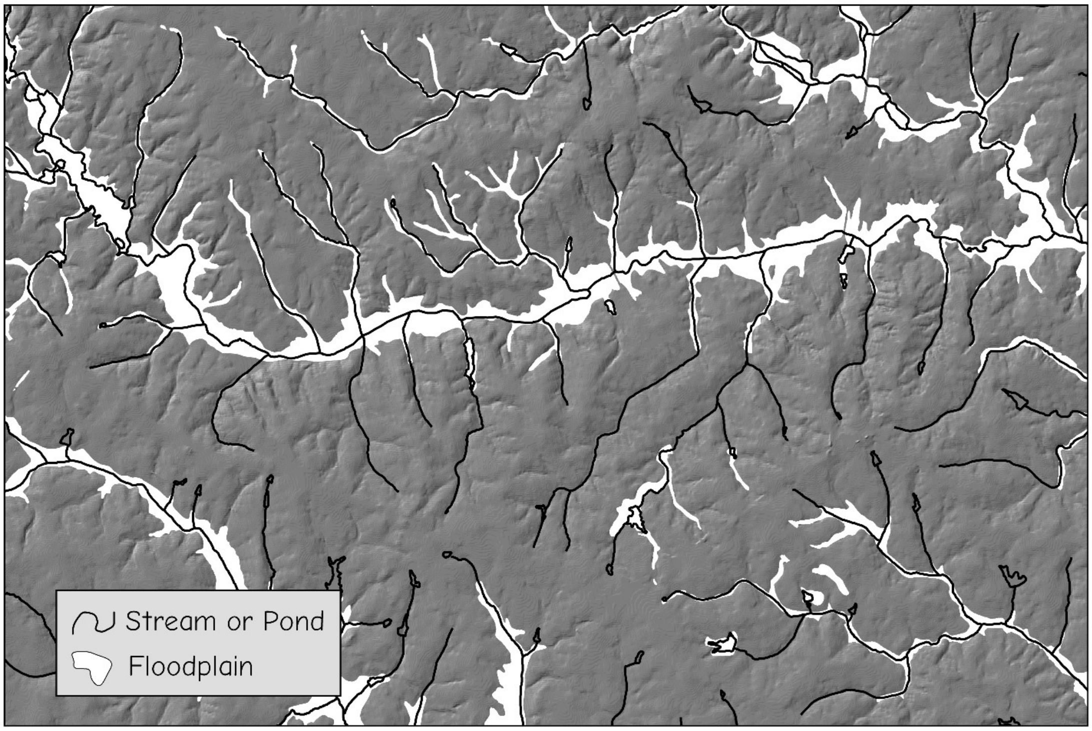


Figure 7-20: An example of FEMA floodplain data for a region near Morgan, Georgia, USA.

Climate, Geology, and Other Environmental Data

Other spatial environmental data sets are available, including climate, water chemistry, energy resources. Here we provide examples, but there are many others.

The National Climatic Data Center (NCDC) maintains historical climate records for the United States, and provides their data through a Web portal (http://gis.ncdc.noaa.gov/maps). Recording stations may be selected by various criteria, including geography, measured variables, or length of record. Climate data have been converted to spatial fields, and are distributed through the PRISM initiative (prism.oregonstate.edu, see Figure 7-21).

Mineral resources data are available from the United States Geological Survey, at http://mrdata.usgs.gov/. These data include maps of basic national geology, as well as spatial and tabular data on specialized themes such as mineral deposits, mines, claims, smelters and other processing facilities, and energy resources.

Spatial data are available for a range of other environmental parameters, including air pollution, pollutant and contaminant distribution, and some water pollutants through the Environmental Protection Agency, many at www.epa.gov/data/.

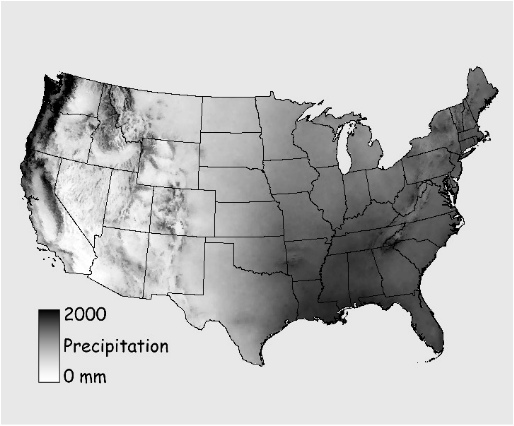


Figure 7-21: U.S. average precipitation, measured from 1971–2000, interpolated from weather stations across the United States to create a raster grid (data from the PRISM project).

Digital Census Data

The United States Census Bureau developed and maintains a database system to support the national census. This system is known as the Census TIGER system (Topologically Integrated Geographic Encoding and Referencing). The TIGER system is used to organize areas by state, county, census tract, and other geographic units for data collection and reporting. It also allows the assignment of individual addresses to geographic entities. The census TIGER system links geographic entities to census statistical data on population size, age, income, health, and other factors (Figure 7-22). These entities are typically polygons defined by roads, streams, political boundaries, or other features. The TIGER system is a key government tool in the collection of census data. TIGER also aids in the application of census data during the apportionment of federal government funds, in congressional redistricting, in transportation management and planning, and in other federal government activities.

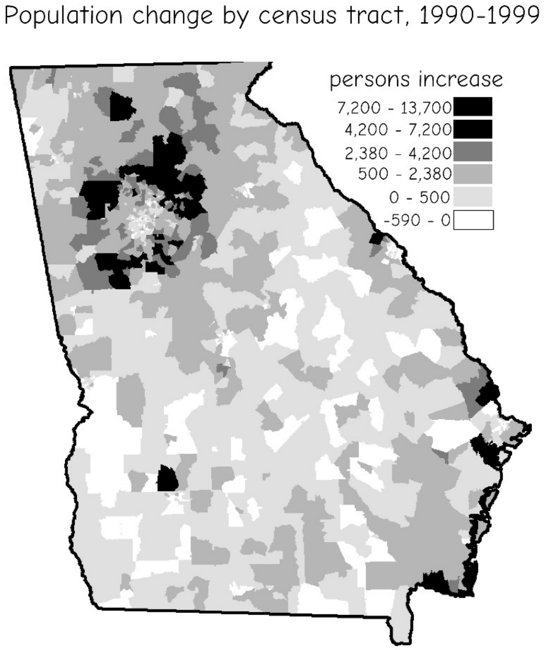


Figure 7-22: Digital census data provide spatially referenced demographic and other data for the U.S.

TIGER/Line files are at the heart of the system. They define line, landmark, and polygon features in a topologically integrated fashion. Lines most often represent roads, hydrography, and political boundaries, although railroads, power lines, and pipelines are also represented. Polygon features include census tabulation areas such as census block groups and tracts, and area landmarks such as parks and cemeteries. Point landmarks such as schools and churches may also be represented. Points, lines, and polygons are used to define these features (Figure 7-23).

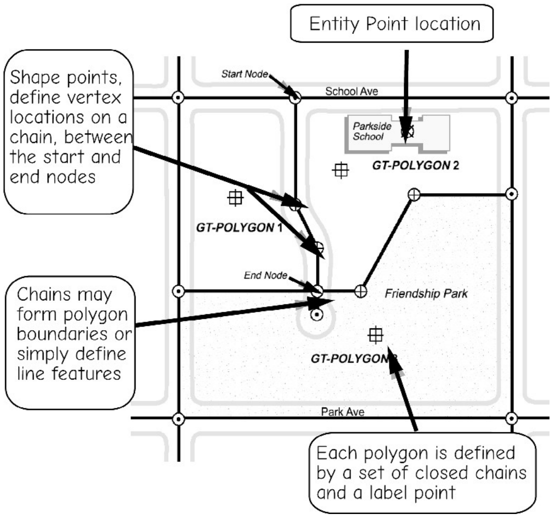


Figure 7-23: TIGER data provide topological encoding of points, lines (chains), and polygons (U.S. Dept. of Commerce).

Nodes and vertices are used to identify line segments. Topological attributes are attached to the nodes and lines, such as the polygons on either side of the line segment, or the line segments that connect to the node. Point landmarks and polygon interior points are other topological elements of TIGER/Line files.

TIGER/Line files contain information to identify street address labels. Starting and ending address numbers are recorded corresponding to starting and ending nodes (Figure 7-24). Addresses may then be assigned within the address range. The system does not allow specific addresses to be assigned to specific buildings. However, it does restrict the addresses on a city block to a limited range of numbers, something of great use to field workers responsible for collecting census information.

TIGER/Line files are organized by different record types. A collection of records reports the location and attribute information

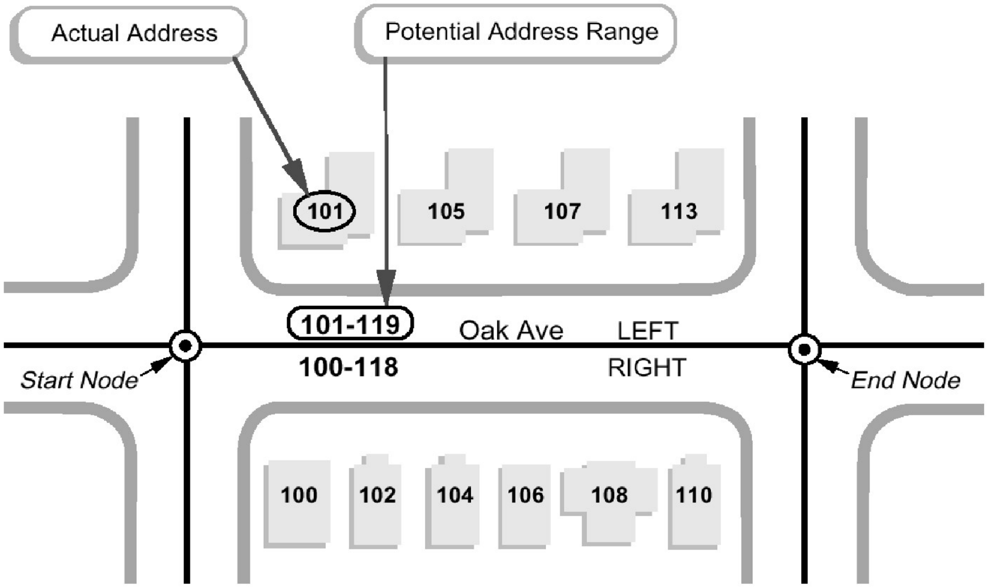


Figure 7-24: TIGER data provide address ranges for line (chain) segments. These ranges may be distributed across the line, giving approximate building locations on a street (U.S. Dept. of Commerce).

about a set of census features, including the location, shape, addresses and other census attributes for a county. There is an identifier based on the United States Federal Information Processing Standards Code (FIPS) that is used to identify the file and record type.

Census data are distributed as ESRI Geodatabases and as shapefiles for various geographic units. In addition, specialized software packages are available to ingest TIGER/Line and related census and other federal government data files to data layers in specific GIS formats. Data are collected on business activities, health, crime, among other topics. These data may be extracted in a customized manner.

Many United States government data sets are provided with codes compatible to United States Census Bureau data. For example, data are delivered in census-compatible units and codes by the United States Department of Education (http://nces.ed.gov/ccd/), the United States Department of Transportation (http://www.bts.gov), and the United States Centers for Disease Control (http://wonder.cdc.gov). Data from these organizations are delivered with codes needed to link statistics to geography, for example, the average traffic fatality rate from 1997 through 2006, as shown in Figure 7-25.

Post-Processing

While many data are available, they often require processing after download to render them useful. This may include some form of editing to winnow table variables or remove unwanted features, or raster resampling, or coordinate projection, or datum transformation, or conversion among data models, all activities covered in previous chapters.

Processing may also entail type conversion, e.g., many point data are distributed as coordinates in a table, and these must be converted to point features in a data layer. Most GIS softwares provide utilities to convert X and Y coordinate data to shapefiles,

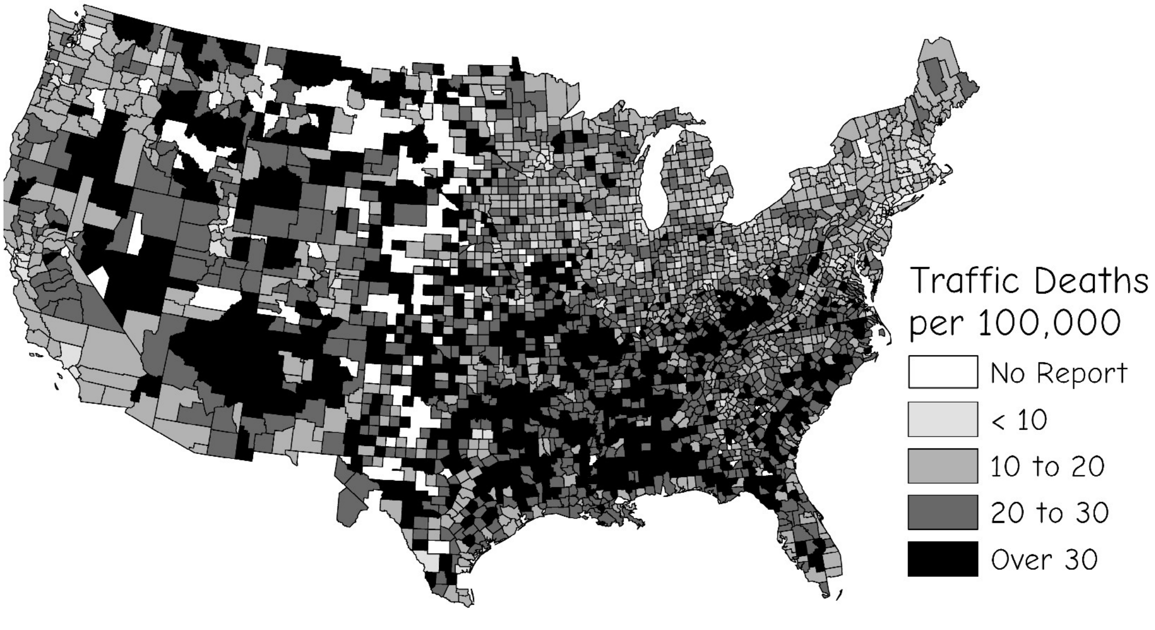


Figure 7-25: Traffic fatality data, showing average number of deaths per 100,000 persons over 1997-2006, derived from data reported by the U.S. Centers for Disease Control. These data are published with links to U.S. Census recognized geographies, as are much other data collected by the federal government. Here, Federal Information Processing Standard (FIPS) codes were published with the CDC data, and used to link to county boundary files. Note the generally high death rates in some southern and interior western counties, and that New York drivers, counter to reputation, appear to be among the safest on the roads.

geodatabase layers, or other common spatial data formats.

Processing may also require conversion among data models, e.g., from vector to raster. Most GIS softwares provide conversion tools among data models, although on conversion we must bear in mind the conceptual differences among models. Remember that vector data don't store information on areas outside of features, for example, the areas outside of polygons are undefined. When converting polygon data to raster data, the raster cells in these "in between" locations may be assigned a value that indicates "null" or "nodata", or otherwise flagged as unknowns (Figure 7-26). These nodata values are sometimes set to a specific numeric value, often zero or an implausible negative number such as -9999, or some other value that is not present in the feature data values. This may affect further processing, and these unknown areas often must be modified before subsequent analysis. Some spatial functions are nonsensical on these codes, e.g., the natural log functions when applied to negative numbers. Other softwares simply won't apply most functions to nodata raster cells, returning an output value of nodata when a cell is encountered in any function or processing. Typically there are functions which explicitly evaluate cells for nodata values, and allow re-assignment. These and other processing tools for vector and raster data are described from Chapter 9 onwards.

Summary

Digital data are available from a number of sources, and provide a means for rapidly and inexpensively populating a GIS database. Most of these data have been produced by government organizations and are available at little or no cost, often via the internet. Data for elevation, transportation, water resources, soils, population, land cover, and imagery are available, and should be evaluated when creating and using a GIS.

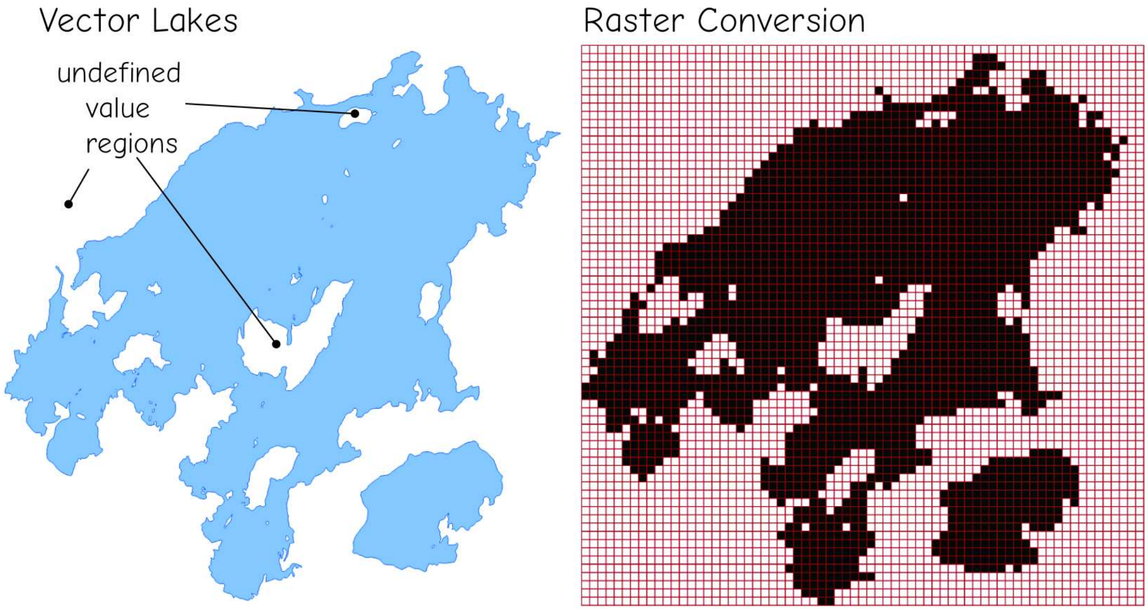


Figure 7-26: Data available for download may be in a non-target data model, here vector polygons that we wish to use in raster processing. A vector model holds no information for areas outside of polygons. Raster models define values everywhere in their extent. "No data" values may be assigned for the undefined regions, which may constrain further processing, until they are changed.

Suggested Reading

Broome, F.R., Meixler, D.B. (1990). The TIGER database structure. *Cartography and Geographic Information Systems*, 17:39–47.

Carter, J.R. (1988). Digital representations of topographic surfaces. *Photogrammetric Engineering and Remote Sensing*, 54:1577–1580.

Decker, D. (2001). *GIS Data Sources*. New York: Wiley.

Di Luzio, M., Arnold, J.G., Srinivasan, R. (2004). Integration of SSURGO maps and soil parameters within a geographic information system and nonpoint source pollution model system. *Journal of Soil and Water Conservation*, 59:123–133.

Duncan, D.T., Aldstadt, J., Whalen, J., Melly, S.J., Gortmaker, S.L. (2011). Validation of the Walk Score for estimating neighborhood walkability: an analysis of four U.S. metropolitan areas. *International Journal of Environmental Research in Public Health*, 8:4160–4179.

Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., Tyler, D. (2002). The national elevation dataset. *Photogrammetric Engineering and Remote Sensing*, 68:5–11.

Goodchild, M. F., Anselin, L., Deichmann, U. (1993). A framework for the areal interpolation of socioeconomic data. *Environment and Planning*, 25:383–397.

Gorokhovich, Y., Voustianiouk, A. (2006). Accuracy assessment of the processed SRTM-based elevation data by CGIAR using field data from USA and Thailand and its relation to the terrain characteristics. *Remote Sensing of Environment*, 104:409–415.

Harvey, F., Leung, Y. (Eds.) (2015). *Advances in Spatial Data Handling and Analysis*. Berlin: Springer.

Horner, C., Huang, C., Yang, L., Wylie, B., Coan, M. (2004). Development of a 2001 national landcover database for the United States. *Photogrammetric Engineering and Remote Sensing*, 70:829–840.

Lytle, D.J., Bliss, N.B., Waltman, S.W. (1996). Interpreting the State Soil Geographic Database (STATSGO). Goodchild, M.F., Steyaert, L.T., Parks, B.O., Johnston, C., Maidment, D., Crane, M., Glendinning, S. (Eds.). *GIS and Environmental Modeling: Progress and Research Issues*. Fort Collins: GIS World.

Marx, R.W. (1986). The TIGER system: automating the geographic structure of the United States Census. *Government Publications Review*, 13:181–201.

Maune, D.F. (2007). *Digital Elevation Model Technologies and Applications: The DEM User's Manual* (2nd ed.). Bethesda: American Society of Photogrammetry and Remote Sensing.

Openshaw, S., Taylor, P. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. N. Wrigle (Ed.). *Statistical Applications in the Spatial Sciences*. London: Pion.

Smith, B., Sandwell, D. (2003). Accuracy and resolution of shuttle radar topography mission data. *Geophysical Research Letters*, 30:1–20

Taylor, P.J., Johnston, R.J. (1979). *Geography of Elections*. Hammondsworth: Penguin.

Wilen, B.O., Bates, K.M. (1995). The US Fish and Wildlife Service's National Wetlands Inventory Project. *Vegetatio*, 118:153-169.

Study Questions

7.1 - What are some advantages and disadvantages of using digital spatial data?

7.2 - What are the most important questions you must ask before using already developed spatial data?

7.3 - For each of the following data sets, tell us who produces them, what are the source materials, what do the data sets contain, their grain sizes and accuracies, and how they are delivered: digital raster graphics (DRGs), digital line graphs (DLGs), digital elevation models (DEMs), digital orthophotoquads (DOQs), digital floodplain data, National Wetlands Inventory data (NWI), SSURGO and STATSGO soils data, TIGER census data, and national land cover data (NLCD) sets.

7.4 - What is edge matching and why is it important?

7.5 - Identify and describe the characteristics of three different sources of digital elevation data. What are the pros and cons of each source?

7.6 - Visit one of the websites mentioned in this chapter, or in the appendices at the end of this book, and download several data layers of an area of interest. If you have access to a GIS, try to import these data and display them.

8 Attribute Data and Tables

Introduction

We have described how spatial data in a GIS are often split into two components, the coordinate information for object geometry, and the attribute information for the nonspatial properties of objects. Because these nonspatial data are frequently presented to the user in tables, they are often referred to as *tabular data*. Tabular data summarize the most important characteristics of each object, for example, attributes about counties (Figure 8-1). In this example, the attributes include the county name, Federal Information Processing Standards (FIPS) code, population, area, and population density.

Attribute information in a GIS is typically entered, analyzed, and reported using a *database management system* (DBMS), a specialized computer program for organizing and manipulating data. The DBMS stores the properties of geographic objects and the relationships among the objects. A DBMS incorporates software tools for managing tabular data, including those for efficient data storage, retrieval, indexing, and reporting. DBMSs were initially developed in the 1960s, and refinements since then have led to robust, sophisticated systems employed by government, businesses, and other organiza-

Name	FIPS	Pop90	Area	PopDn
Whatcom	53073	128	2170	59
Skagit	53057	80	1765	45
Clallam	53009	56	1779	32
Snohomish	53061	466	2102	222
Island	53029	60	231	261
Jefferson	53031	20	1773	11
Kitsap	53035	190	391	485
King	53033	1507	2164	696
Mason	53045	38	904	42
Grays Harbor	53027	64	1917	33
Pierce	53053	586	1651	355
Thurston	53067	161	698	231
Pacific	53049	19	945	20
Lewis	53041	59	2479	24

Figure 8-1: Data in a GIS include both spatial (left) and attribute (right) components.

tions. A somewhat standard set of DBMS tools and methods have been developed and are provided by many vendors.

Note that the terms DBMS and database are sometimes used interchangeably. In most cases this is incorrect and in all cases imprecise. A DBMS is a computer program that allows you to work with data. A database is an organized collection of data, often created or manipulated with the help of a DBMS. The database may have a specific form dictated by a DBMS, but it is not the system.

Students often struggle with relational databases at first, and often ask, “Why bother? Can’t we just use a spreadsheet?” Many more people are familiar with spreadsheet forms, programs, and manipulation, and don’t see the value added when adopting a DBMS. A short example may help explain their value.

Consider the file shown in Figure 8-2, representing business orders. Each row records the purchaser, an order number, and the items ordered. Spreadsheets typically present data like this in a single, “flat file” with two dimensions, rows and columns. Because orders may contain multiple items, we need multiple columns with copies of the item/quantity pair. For example, order number five by Atom Ant includes two items, two B52s and two CR7s, while order number three by Paul Smith has four items, or a total of 8 columns for items. Larger orders would require additional columns.

This storage form has two characteristics. First, we either have to limit the number of items per order (rarely a good thing for a business), or else not know how many columns our database might have, which would complicate programing and management. Second, and more importantly, we can easily have most of the storage in our database contain nothing. We may have thousands of orders with one or two items. However, if we have one order with 50 different items, we have to add enough columns to accommodate 50 item/quantity pairs, even in orders with one or two items. As with orders 1, 2, 4, 6, 7, and 8 in the table below, many of the cells will be empty. We can easily have a database that is mostly empty cells. Computer memory has become quite inexpensive, so you may think this a minor disadvantage. Still, more data means longer processing times, to the extent that the database may not function. This flat file structure is flawed, in both inefficient use of space and slow processing.

There is another obvious disadvantage of this structure. Note that there are two orders from Paul Smith, order number 3 and order number 8. This means there are redundant sets of information for the same customer. We have his first and last name, address, and phone number repeated in both

name	surname	address	phone #	order #	item	qty	item	qty	item	qty	item	qty
Leo	Durocher	112 Beal St	5-1307	1	CR7	1						
Rudy	Valentini	1 Hispanola Dr	4-2706	2	F15	1						
Paul	Smith	99 Upstate Ln	0-0000	3	GTO	3	F15	1	B52	1	SR71	1
Adam	Smith	1 Wall St	1-2334	4	626	1						
Atom	Ant	685 Hanbar Rd	4-1222	5	B52	2	CR7	2				
William	Smith	202 Dinkytown	9-9199	6	F111	2						
Alice	Paul	5 Free St.	4-4178	7	SR71	1						
Paul	Smith	99 Upstate Ln	0-0000	8	F15	1						

Figure 8-2: An example of database, in a flat file format.

orders. This wastes space and makes editing more cumbersome and error prone. If Paul Smith changes his phone number, we must search through every line in our database and change every instance of an order that contains Paul Smith's phone number. Redundant storage and editing is an additional disadvantage of a spreadsheet-like file system.

Functions and programs may be written to address the inefficient use of space, slow processing, and difficulty editing in spreadsheets or other apparently flat file formats. These programs often require specific knowledge of the spreadsheet structure, and so depend on the arrangement and number of columns. While these workarounds are possible, they are often complicated and require substantial program maintenance. Database management systems were developed to overcome these redundancies and inefficiencies by adding structure to data files in a standard way. While spreadsheets may be used for simple collections of data, DBMSs are better for most applications that process large amounts of data.

DBMSs provide other advantages. They may provide *data independence*, a valuable characteristic when working with large data sets. Data independence allows us to make changes in the database structure in ways that are transparent to any use or program. This means restructuring the database does not require a user or programmer to modify their procedures. Before data independence became widespread, organizations frequently spent considerable time and money rewriting applications and retraining users with each change of the data structure. Data independence avoids this.

DBMS may also provide for *multiple user views*. Different users may require different information from the database, or the same information delivered in different formats or arrangements. Profiles can be developed that change the way data are provided to each program or user. DBMSs are able to automatically reformat data to meet the viewing preferences. The DBMS eliminates the need to have copies of the data for each user, by changing the presentation to meet each specific need.

A DBMS also allows *centralized control and maintenance* of important data. One "standard" copy of the data may be maintained and updated on a regular, known basis. These data may be time stamped or provided with a version number to aid in management. These data are then distributed to the various users. A single person or group may be charged with maintaining data currency, quality, and completeness, and with resolving contradictions or differences among various versions of the database.

Adopting a DBMS may come at some cost. Specialized training may be required to develop, use, and maintain a database. Defining the components of a database and relationships among them may be a complex task that may require specialists. Structuring the database for efficient access or creating customized forms will often require significant effort. The software itself may be quite expensive, although free, stable, open source database management software is available. Users may need to optimize speed for certain operations that are too slow in a DBMS. However, for many users, the value of the DBMS and database development far outweighs these costs.

Database Components and Characteristics

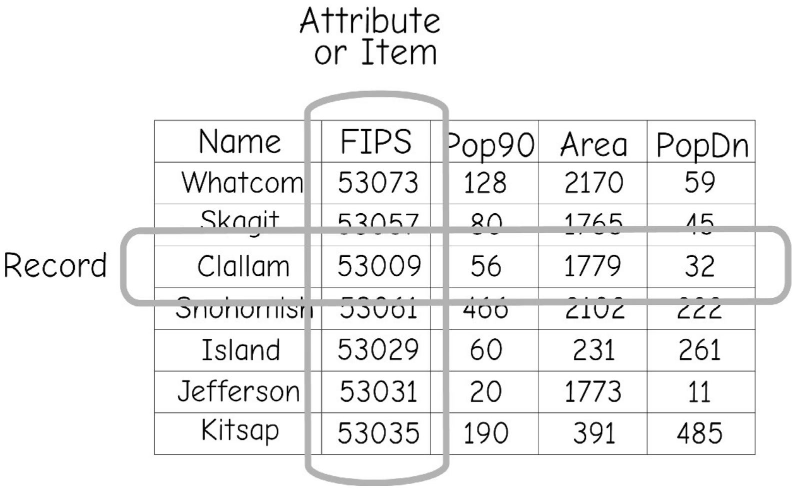
The basic components of a traditional database are *data items* or *attributes*, the indivisible named units of data (Figure 8-3). These items can be identifiers, sizes, areas, coordinates, colors, or any other suitable characteristic used to describe things. Attributes may be simple, for example, one word or number, or they may be compound, for example, an address data item that consists of a house number, a street name, a city, and a zip code.

Items have a *type* and a *domain* that restrict the values they may take. Types define essential characteristics of an item. Common types include real numbers, integer numbers, both of various lengths, hexadecimal numbers, text fields, hyperlinks, and binary large objects (blobs). Domains define the acceptable values an item may take, for example, integers may be restricted to be larger than 0 but smaller than 10, or there may be a type name "color" that can only take on the values "red", "green", "blue", "yellow", "cyan", or "magenta."

A collection of related data items that are treated as a unit represents an *entity*. In a GIS, the database entities are typically roads, counties, lakes, or other types of geographic features. A specific entity, such as a specific county, is an *instance* of that entity. Entities are defined by a set of attributes and associated geographic data. In our example in Figure 8-3, the attributes that describe a county include the name, a FIPS code, the 1990 population in thousands of persons, area, and the population density. These related data items are often organized as a row or line in a table, called a *record*. A *file* may then contain a collection of records, and a group of files may define the database. Specific database systems often define the terms differently for each of these parts. For example, in the relational database model, the record may be called a *row* or an *n-tuple*, and the tables referred to as a *relational table*, or sometimes as just a *relation*.

You should note that the concept of an entity, when referred to in a database, may be slightly different than an entity in a GIS data model. This difference stems from two different groups, geographers and computer scientists, using a word for different but related concepts. An entity in a geographic data model is often used for the real-world thing we are trying to represent with a cartographic object. These entities are typically a physical phenomenon, e.g., a lake, city, or building, but they may also be a conceptual phenomenon, such as a property boundary. In contrast, computer scientists and database managers often define an entity as the principal data object about which information will be collected. In the DBMS literature, the entity is the data object that denotes a physical thing, and not the thing itself. Thus, properties of entities and relationships

Figure 8-3: Components of an attribute data table.



among entities refer to the structure of the DBMS. This is a subtle distinction in terminology, but these different definitions can lead to confusion unless the difference in meanings is noted. For the remainder of this chapter, we will use the definition of an entity as a data object.

A DBMS typically supports complex structures, primarily to provide data security, to maintain stability, and to allow multiple users to access the same data simultaneously. Database users often demand shared access, when multiple users or programs can open, view, or modify a data set simultaneously. If each program or user has direct file access, multiple copies of a database may be open for modification at the same time (Figure 8-4, top). Multiple users may try to write to the data file simultaneously, with unforeseen results. The data saved may be the most recent, the first updates, or some mix in between. Because of these hazards with direct file access, a DBMS may be designed to manage multiuser access (Figure 8-4, bottom). Some DBMSs manage shared files and data, and enforce a predetermined precedence in simultaneously accessed files. The DBMS may act as an intermediary between the files and the application programs or user. The DBMS may prevent errors due to simultaneous access. Other DBMS programs do not manage simultaneous access, and users of such systems generally must avoid opening multiple copies of the database at one time.

The DBMS is sometimes referred to as a database *server*, and the applications programs as *clients*. The server provides or "serves up" data to the client applications. Clients may be built in by the DBMS vendor, written by the DBMS user, or sold as add-ons by third-party software developers. These clients may operate on the same computer as the DBMS server software, or they may provide requests from remote machines over a network connection. A single server may be configured to respond to many client programs running on separate machines.

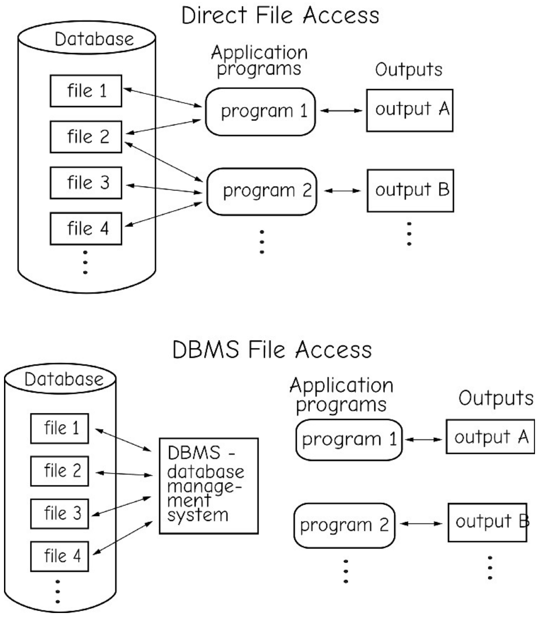


Figure 8-4: Direct and database management system file access (adapted from Aronoff, 1991).

The separation of data and functions into multiple levels is often referred to as a *multi-tiered architecture* (Figure 8-5). Data are primarily stored at the lowest tier. These data may be of diverse types, including coordinate data, attributes, text, images, sound, video recordings, or other important, persistent data.

Data sets at the lowest tier may be managed by an individual database system (Figure 8-5). The system or programs that access the first tier, at the bottom of a multi-tiered system, is often called a *transaction manager*. This transaction manager typically takes requests from higher tiers and searches the relevant portions of the database to identify the requested data, or perform the requested operation.

The next tier in the multi-tiered architecture is often referred to as an *applications server* (Figure 8-5). The use of the term server may be a bit confusing, because server is also used to describe a computer on which data are stored, and also to describe the database management program in two-

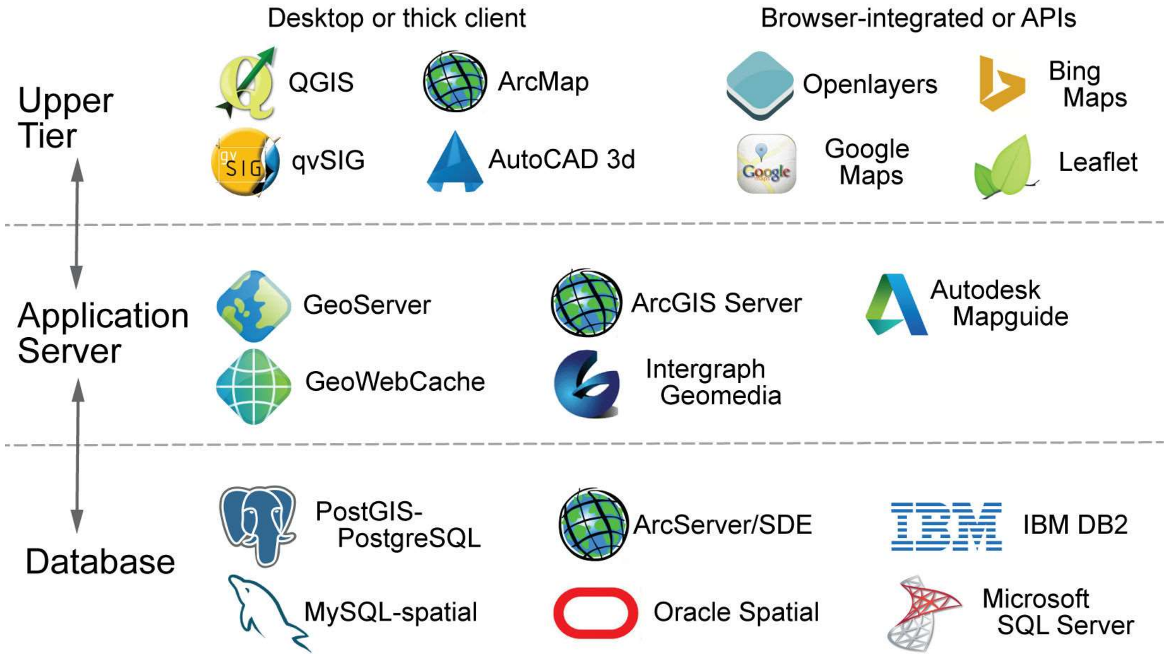


Figure 8-5: Multi-tiered architecture, and common software alternatives. Data are stored and accessed from databases on the bottom tier, and data viewed requests made at the upper tier. Applications servers act upon the request and translate/process information passed between upper and bottom tiers. In well-behaved systems that adhere to defined standards, the components may be mixed, e.g., you may access data in ArcSDE from QGIS, or data stored in PostGIS from ArcMap (adapted from Boundless).

tier systems. In the context of a multi-tiered database architecture, an *applications server* is software that passes requests from higher-level tiers to the transaction manager. It converts input from above into a set of instructions the transaction manager below can "understand." A single request from a higher tier may require the applications server to query several databases. For example, a real estate agent may want to identify all houses in a certain price range, in good school districts, and near rapid transit stations for a prospective buyer. The applications server may generate three different requests — one to identify houses in a price range, a second to identify good school districts, and a third to find rapid transit stations. The applications server may then perform the operations to determine where these important criteria are met.

Besides passing requests to lower tiers, the application server may also perform other tasks. For example, it may determine if the real estate agent has the proper clearance to access the housing data. In addition, the applications server may check to see if the agent has a profile, and so handle requests in a certain way.

The uppermost tier of multi-tiered architectures is typically a *user interface* (Figure 8-5). This tier may be a display by a single-purpose or topic-specific program such as a GIS, or a Web-based interface with the primary purpose of gathering requests from the user, and presenting information back to the user based on those requests.

Multi-tiered architectures are adopted primarily to insulate the user interface from the processing and data at lower tiers, and to allow access to a more diverse range of data through the lowest tiers. The parts are easier to change when they are isolated, and new, different resources may be more easily integrated. If a company decides to redesign their data entry interface, they may do so easily if the user interface is distinct from the tiers below. They do not have to worry about how the applications server or transaction manager access the databases lower down. The integration of a new database technol-

ogy is often easier with multi-tiered architectures.

Multi-tiered architectures are by their nature more complex and variable than two-tier architectures. For example, different implementations of multi-tiered architectures may split operations differently between the tiers. In our real estate example above, one architecture may incorporate all the database query and processing operations into the transaction manager. Another multi-tiered architecture may perform the query with the transaction manager, and perform the spatial operations (houses in good school districts, and good houses near transit stations) in the applications server tier. Neither architecture is universally better; rather, an organization must adopt an architecture that best suits its needs.

Multiuser access adds substantial overhead and complexity to processing. For example, the server must ensure that when several copies of a database are accessed, changes to the database must be reconciled on resubmission. If two different clients have altered different variables in a database, these two sets of changes must be integrated when the database is stored. If the updates from two clients conflict, such as when one client deletes a record while a second client modifies a value for the same record, the program must resolve the differences; perhaps one user has higher priority, the most recent changes are enforced, or a message is sent to an operator noting the ambiguity.

Physical, Logical, and Conceptual Structures

A database may be viewed as having conceptual, logical, and physical structures. These structures define the entities and their relationships, and specify how the data files or tables are related one to another.

The conceptual structure is often represented in a *schema.* A schema succinctly describes the database structure in standard shorthand notations, usually via *entity-relationship* diagrams, also known as E-R diagrams. We will not describe E-R diagrams or other conceptual methods here, as they are more appropriate for a more advanced course.

A schema is a compact graphical representation of the conceptual model, the entities, and the relationships among them. The relationships may be one-to-one, between one entity and another, say from a row representing a purchase order in a database table to a row representing customer in another table. Connections between tables may be one-to-one, from one customer to a single order, or one-to-many, from a customer to several orders. These relationships may be represented on figures by lines connecting the entities.

The actual connections within the computer files may by achieved in many ways. One common method uses file pointers to connect records in one file with those in other files. Much of this structure is designed to speed access, aid updates, and provide data integrity. This structuring is part of the *physical design* of the database. The design typically strives to physically cluster or link data used together in processes so that these processes may be performed quickly and efficiently.

Relational Databases

Relational databases have grown to become the most common database design since their introduction in 1968, for various reasons. The relational model is more flexible than most other designs. The tables structure does not restrict processing or queries, and the organization is not too difficult to understand, learn, and implement relative to other database designs. It can accommodate a wide range of data types, and it is not necessary to know in advance the kind of queries, sorting, and searching that will be performed on the database.

There are typically a cluster of tables, or relations, in a relational database design, as

shown for forest and related recreation data in Figure 8-6. Entities are represented by rows in a table. In our forest data example, there may be a forest table with a row for each forest, and other tables representing the trails, trail features, and recreational opportunities. As noted earlier, the rows are also called *records* or *tuples*.

Tables are related through *keys*, one or more columns that meet certain requirements and may be used to index the rows. Keys are often a column that uniquely identifies every row in a table. We often assign a unique number or code to be a key, for example, a Social Security number may be used as a key for a set of people in the United States. No two people have the same valid Social Security number, so we can use the number to connect a row of information to a specific person.

Keys are used to join data from one table to associated data in another table (Figure 8-7). Keys are the "key" to the utility and flexibility of relational databases. They allow us to mix and match data from various tables; to display data differently for different projects or audiences; to organize our data in ways that help us more quickly search, select, and update our data; and to isolate our data from calling programs or changes in computer hardware.

Figure 8-7 shows a join of our forest and trails data in a relational data structure, through the key `Forest-ID` in the `Forests` and `Trails` tables. This shows how keys allow us to break our data up into several tables and reap all the benefits described above, while providing a mechanism to link between tables. Note that this linkage of separate tables is usually transparent to the end user, in that all or subsets of the forest and trails data in Figure 8-7 may be displayed on

Forests

Forest Name	Forest-ID	Location	Size
Nantahala	1	N. Carolina	184,447
Cherokee	2	N. Carolina	92,271

Trails

Trail Name	Forest-ID
Bryson's Knob	1
Slickrock Falls	2
North Fork	1
Cade's Cove	1
Cade's Cove	2
Appalachian	1
Appalachian	2

Characteristics

Trail Name	Feature	Difficulty
Bryson's Knob	Vista	E,M
Bryson's Knob	Ogrth	E,M
Slickrock Falls	Ogrth	M
Slickrock Falls	Wfall	M
North Fork	-	M
Cade's Cove	Ogrth	E
Cade's Cove	Wlife	E
Appalachian	Wfall	M,D
Appalachian	Ogrth	M,D
Appalachian	Vista	M,D
Appalachian	Wlife	M,D
Appalachian	Cmp	M,D

Recreational features

Feature	Description	Activity 1	Activity 2
Wfall	Waterfall	Photography	Swimming
Ogrth	Old-Growth Forest	Photography	Hiking
Vista	Scenic Overlook	Photography	Viewing
Wlife	Wildlife Viewing	Photography	Birding
Cmp	Camping	Camping	-

Figure 8-6: Forest data in a relational database structure.

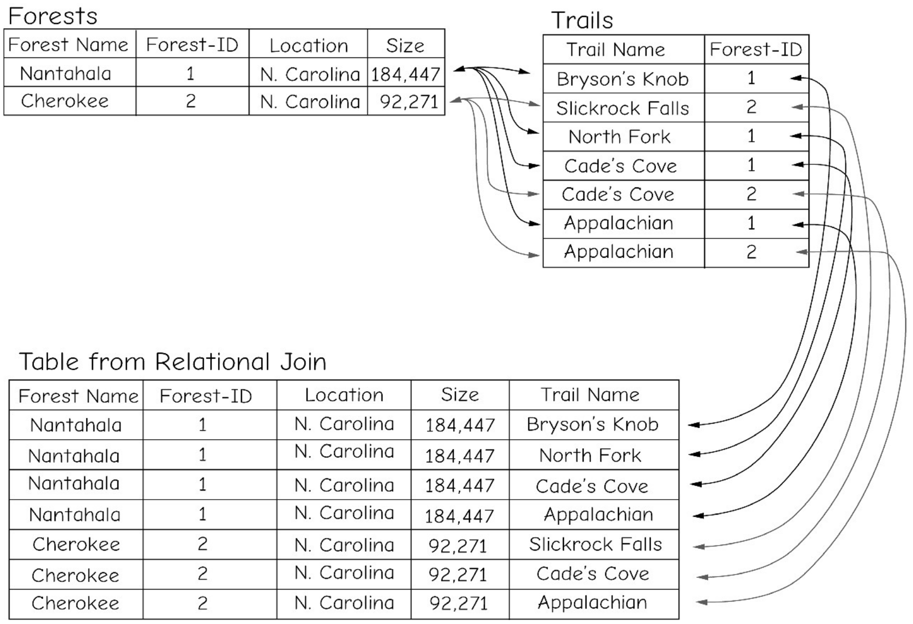


Figure 8-7: Forest and trails data in a relational data structure. Rows hold records associated with an entity, and columns hold items. A key, here Forest-ID, is used to join tables.

screen or printed as one continuous table. Data from three or more distinct tables in a DBMS are often joined, and columns subsets displayed in what appears to be one table to a user. The joins and columns may be changed, depending on the user.

The general definition of relational databases defines a *relational algebra*. This relational algebra takes relations (tables) as input and returns relations as output. The algebra combines or splits tables, either by rows or columns, to generate subset or expanded tables. The relational algebra may also be used to specify constraints, requirements, and security on a database.

Given their importance, there are some restrictions on keys. For example, null values are typically not allowed to be part of a key. There may be many potential keys (columns that uniquely identify each row), but typically one is chosen for use, called a *primary key*. Most tables in the database will have a primary key, and keys in a table are frequently used to combine tables. Some keys are used to index and add flexibility in selecting data. Too few keys may result in difficulties searching or sorting the database.

Primary Operators

The relational algebra supports eight primary operators: *restrict*, *project*, *union*, *intersection*, *difference*, *product* (all combinations of a given set of variables recorded in a database), *join* (combine tables based on matching attribute values), and *divide* (facilitate queries based on condition). These operations are applied in queries to select specific records and items. These operations are depicted graphically in Figure 8-8 and Figure 8-9.

Restrict and project operations select based on rows and columns, respectively, to provide reduced tables. Restrict, also known and described as a *table query*, serves up records based on values for given variables. The restrict in Figure 8-8a is specified to restrict the current set to those that have a size that is big or huge—all other entries in the relation are not selected (remember, tables are called "relations" in a relational database). The restrict then only returns four of the seven records, as shown in Figure 8-8a.

a) restrict

ID	type	color	size	age
1	a	blue	big	old
2	c	green	big	young
3	a	red	small	mid
4	d	black	big	older
5	x	mauve	tiny	oldest
6	g	dun	huge	young
7	c	ecru	small	mid

restrict →

ID	type	color	size	age
1	a	blue	big	old
4	d	black	big	older
6	g	dun	huge	young
2	c	green	big	young

b) project

ID	type	color	size	age
1	a	blue	big	old
2	c	green	big	young
3	a	red	small	mid
4	d	black	big	older
5	x	mauve	tiny	oldest
6	g	dun	huge	young
7	c	ecru	small	mid

project →

ID	color	size
1	blue	big
2	green	big
3	red	small
4	black	big
5	mauve	tiny
6	dun	huge
7	ecru	small

c) product

No.	Dir.
1	N
2	S

product

App.
Yes
No

→

No.	Dir.	App.
1	N	Yes
2	S	Yes
1	N	No
2	S	No

d) divide

type
m
n
r

divide by

size
1
2

per

type	size
m	1
m	2
m	3
m	4
n	2
r	1
r	3

→

type
m

Figure 8-8: Relational algebra as originally defined supported eight operators. The first four are shown here: restrict, project, product, and division. The remaining four are shown in the next figure (modified from C.J. Date, 2004).

Restrict operations can be compound and complex and involve more than one attribute. Restrict operations most often return a reduced set of rows for a table as output. Examples of more complex restrict operations, or table queries, will be shown later in this chapter.

Project operations return entire columns for a table, in effect subsetting the table vertically, as shown in Figure 8-8b. Database tables may be quite large, and contain hundreds of items. A given analysis may concern only a few of those items, and so the project operation allows only those columns of interest from the table to be subset. This may substantially increase processing speed, reduce the storage space required, and ease viewing and analysis. In the example shown, ID, color, and size are selected from a base relation to create a new relation.

Product operations combine all rows in one table with rows in another table to output a larger table (Figure 8-8c). If there are n rows in the first table, and m in the second, then there are n x m rows in the product. The product defines the complete set of possible combinations that may occur when combining two relations.

The divide operation is often the most confusing of the eight original relational operators, at least to new users. The divide operation is not obviously related to a mathematical divide, at least until the user has accrued some experience using the relational divide. The relational divide is analogous to the mathematical divide in that there is a target (the dividend) that is divided by another table (divisor). However, the confusion comes in that this is done relative, or per, a third table.

Divide operations are generally used in queries that would use the word “all” in a natural language description of the request, for example, “get states with all three major ethnic groups.” A dividend is divided by a divisor over a table. For example, the state list is divided by the ethnic group list over a table containing state and ethnic group items. Only those states with all three ethnic groups are returned.

An example of a relational divide operation is shown in Figure 8-8d. The first table with the item named type is divided by the table with the item named size over the table with type and size. This returns a table with the item type and one entry, m. Only those values of type that have records with all values of size are returned. Note that in Figure 8-8d the n is missing size 1,3, and 4, and r is missing size 2 and 4.

The remaining four primary relational operators, as defined by E.F. Codd, typically return records based on membership in two or more tables. These operators are union, intersection, difference, and join, and they are illustrated in Figure 8-9, and described in turn in the following paragraphs.

Before we describe the union, intersect, and difference operations, we must note a limitation in their application. The tables used in these three relational operations must be of the same kind. That means they must have the same set of variables or items. It makes little sense to find the intersection of two tables when they do not share the same set of items, for example, a table of home addresses and a table of plant species. These tables will always have an empty intersection set. There are similar problems when the union and difference relational operations are performed on tables that do not have the same set of items. Therefore, these operations are only defined and allowed in the context of tables of the same kind—that is, tables that have exactly the same items, defined in exactly the same way. This doesn’t mean the tables are identical—there will be different values in the various columns. Rather, the columns for both tables must be of the same data type (e.g., integer, text, real), and all columns must be present in both tables.

A union operation combines tables to return records found in either or both tables. As shown in Figure 8-9a, the tables are “stacked” to return a new table with

a) union

ID	type	color	size	age
1	a	blue	big	old
6	g	dun	huge	young

ID	type	color	size	age
2	c	green	big	young
4	d	black	big	older

union →

ID	type	color	size	age
1	a	blue	big	old
4	d	black	big	older
6	g	dun	huge	young
2	c	green	big	young

b) intersect

ID	color	size
1	blue	big
2	green	big
3	red	small
4	black	big
5	mauve	tiny
6	dun	huge
7	ecru	small

ID	color	size
1	blue	big
5	mauve	tiny
9	ivory	big

intersect →

ID	color	size
1	blue	big
5	mauve	tiny

c) difference

ID	color	size
1	blue	big
2	green	big
3	red	small
4	black	big
5	mauve	tiny
6	dun	huge
7	ecru	small

ID	color	size
1	blue	big
5	mauve	tiny
9	ivory	big

difference →

ID	color	size
2	green	big
3	red	small
4	black	big
6	dun	huge
7	ecru	small

d) join

ID	type
1	a
2	b
3	b
4	a

type	color	size	age
a	blue	big	old
b	dun	tiny	old

join →

ID	type	color	size	age
1	a	blue	big	old
2	b	dun	tiny	old
3	b	dun	tiny	old
4	a	blue	big	old

Figure 8-9: Four of the eight relational algebra operators as originally defined by E.F. Codd: union, intersect, difference, and join.

members of both, but it does not show duplicate records for those entries that appear in both tables. As such, the result of a union is at least the size of the largest of the two tables, and no larger than the sum of the two tables.

The intersection operation returns records that occur in both input tables, and omits records found in only one of the two input tables (Figure 8-9b). Note that the records with ID values of 1 and 5 are the only two that are found in both tables, and so they are the only two included in the output table.

The difference operation returns those records that are in the first, but not the second table (Figure 8-9c). This example shows all those records that are not in both tables, in our example 2 through 4, plus 6 and 7.

The order of table input in a union or intersection operation does not change output. With the difference operator, order usually matters. The set of records returned from the difference of the first table from the second, shown in Figure 8-9c would change if the tables were reversed in Figure 8-9c.

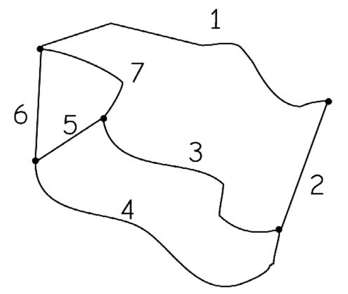
A join operation combines two tables through keys. Values in one or more keys are matched across tables, and the information is combined based on the matching. Figure 8-9d shows an example of a join across two tables, in this case joined through the type item. Each type entry in the table on the left is matched to the type value in the center table, and the data are then joined or related through the values of type. The output records to the right of Figure 8-9d are the combined attributes of both tables. Records in the output table with ID values equal to 1 and 4 have type values equal to a as well as the color, size, and age associated with type a. Those records with type b have the appropriate IDs (2, 3) and the color, size, and age associated with type b.

Hybrid Database Designs in GIS

Data in a GIS are often stored using hybrid designs. Hybrid designs store coordinate data using specialized database structures, and attribute data in a relational database. Thousands to millions of coordinate pairs are often required to represent the location and shape of objects in a GIS. Even with modern computers, the retrieval of coordinate data stored in a relational database design is often too slow. Therefore, the coordinate data are frequently stored using structures designed for rapid retrieval. This involves grouping coordinates for cartographic objects, for example, storing ordered lists of coordinate pairs to define lines, and indexing or grouping lines to identify polygons. Pointers are used to link related lines or polygons, and unique identifiers link the geographic features (points, lines, or polygons) to corresponding attribute data (Figure 8-10).

Topological relationships may be explicitly encoded to improve analyses or to increase access speed. Addresses to the previous and next data are explicitly stored in an indexing table, and pointers are used to connect coordinate strings. Explicitly recording the topological elements of all geographic objects in a data layer may improve geographic manipulations, including determination of adjacencies, line intersection, polygon overlay, and network definition. Coordinates for a given feature or part of a feature may be grouped and these groups indexed to speed manipulation or display.

Hybrid data designs typically store attribute data in a DBMS. These data are linked to the geographic data through unique identifiers or labels that are an attribute in the DBMS. Data may be stored in a manner that facilitates the use of more than one brand of DBMS, and allows easy transport of data from one DBMS to another.



Attribute data

Arc-ID	last-arc	next-arc	type
1	6	2	A
2	1	4	C
3	2	5	F
4	2	5	D
5	6	7	A
6	5	1	F
7	3	6	B

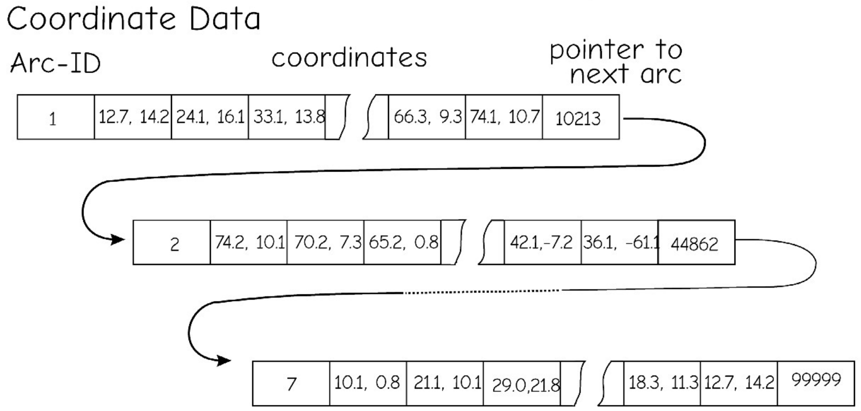


Figure 8-10: A small example of a hybrid database system for spatial data. Attribute data are stored in a relational table, while coordinate data are stored in a network or other structure.

Selection Based on Attributes

The Restrict Operator: Table Queries

Queries are among the most common operations in a DBMS. A query may be viewed as the selection of a subset of records based on the values of specified attributes. Queries may be simple, using one variable, or they may be compound, using one or more conditions on more than one variable. One might search for all the parcels with unpaid taxes, all census blocks larger than a square mile and with at least 200 inhabitants, or all fire hydrants that haven't been pressure tested, are near high-rise buildings, and are farther than 300 m from the nearest other fire hydrant. In concept, queries are quite simple, but basic query operations may be combined to produce quite complex selections.

Many GIS softwares provide a query builder, a graphical user interface (GUI) that helps in applying selection operations (Figure 8-11). Most GUIs include a list of available fields, operations, and a sample or complete display of values for selected fields. The user constructs queries by alternately selecting item names, operations, and entering values. This query may then be applied, and features selected. Often you may save complicated or long expressions, to be reused later on different data sets.

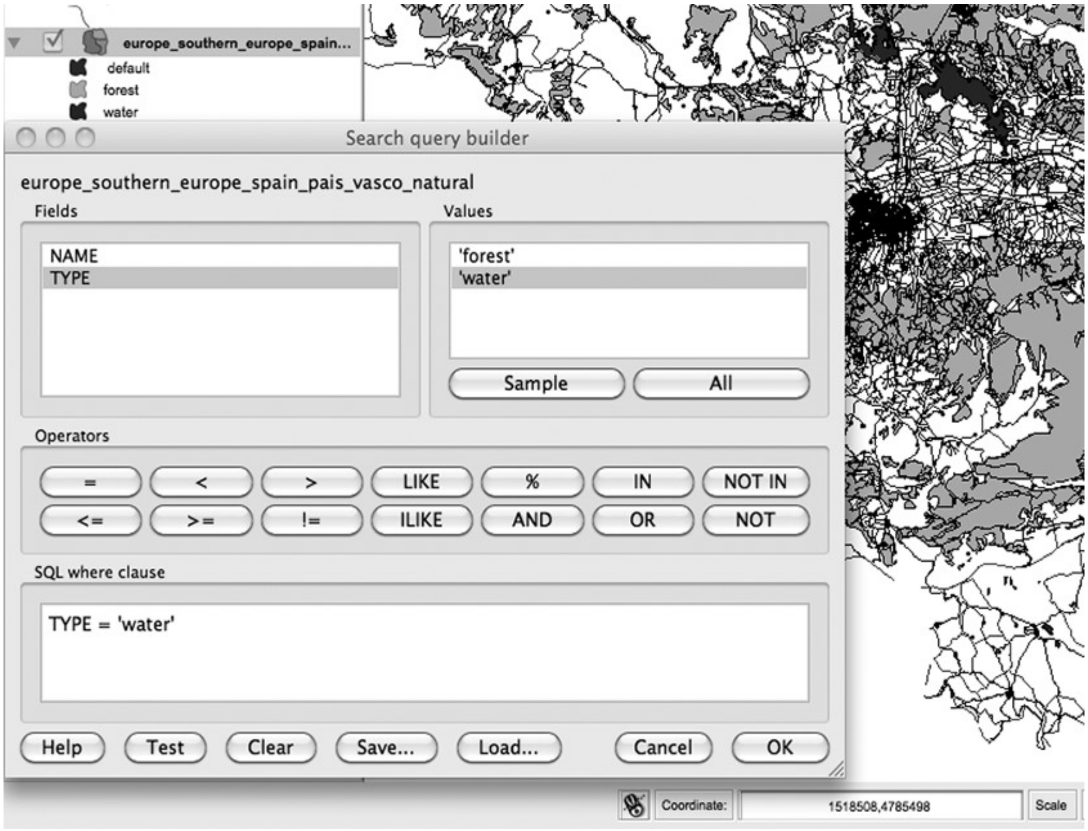


Figure 8-11: A query building GUI of the sort often provided in GIS software, here from QGIS. Selection expressions may be built in the bottom panel by clicking on fields, operators, and values in the upper panels.

The left side of Figure 8-12 demonstrates a simple query. A single condition is specified, Area > 20. The set of selected records is empty at the start of the query. Each record is inspected and added to the selected set if the attribute Area meets the specified criteria. Each record in the selected set is shown in gray in Figure 8-12.

The right side of Figure 8-12 demonstrates a compound query based on two attributes. This query uses the AND condition to select records that meet two criteria. Records are selected that have a Landuse value equal to Urban, and a Municip value equal to City. All records that meet both of these requirements are placed in the selected set, and records that fail to comply with these requirements are in the unselected set. The Boolean operations AND, OR, and NOT may be applied in combination to select records that meet multiple criteria.

AND combinations typically reduce the size of the selected set when compared to the individual component criteria. They provide a more strenuous set of conditions that must be met for selection. In the example on the right side of Figure 8-12, the record with ID = 7 meets the first criterion, Landuse = Urban, but it does not meet the second criterion specified in the AND, Municipality = City. Thus, the record with ID = 7 is not selected. ANDs add restrictions that winnow the selected set.

OR combinations typically increase or add to a selected set in compound queries. OR conditions may be considered as inclusive criteria. The OR adds records that meet a criterion to a set of records defined by previous criteria. In the query on the left side of Figure 8-13, the first criterion, Area > 20, results in the selection of records 2, 4, 5, and 6. The OR condition adds any records that satisfy the criterion Municip = City, in this case the record with ID = 1.

The NOT is the negation operation, and may be interpreted as meaning "select those records that do not meet the condition." The right side of Figure 8-13 demonstrates the negation operation. The operation may be viewed as first substituting equals for the NOT, and identifying all records. Then the remaining records are placed in the selected set, and the identified records placed in the unselected set.

ANDs, ORs, and NOTs can have complex effects when used in compound conditions, and the order or precedence is important in the query. Combinations of these three oper-

Simple selection:

records with Area > 20.0

ID	Area	Landuse	Municip
1	10.5	Urban	City
2	330.3	Farm	County
3	2.4	Suburban	Township
4	96.0	Suburban	County
5	22.1	Urban	City
6	30.2	Farm	Township
7	4.4	Urban	County

AND selection:

records with (Landuse = Urban) AND (Municip = City)

ID	Area	Landuse	Municip
1	10.5	Urban	City
2	330.3	Farm	County
3	2.4	Suburban	Township
4	96.0	Suburban	County
5	22.1	Urban	City
6	30.2	Farm	Township
7	4.4	Urban	County

Figure 8-12: Simple selection, applying one criterion to select records (left), and compound selection, applying multiple requirements (right).

OR selection:
records with (Area > 20.0)
OR (Municip = City)

ID	Area	Landuse	Municip
1	10.5	Urban	City
2	330.3	Farm	County
3	2.4	Suburban	Township
4	96.0	Suburban	County
5	22.1	Urban	City
6	30.2	Farm	Township
7	4.4	Urban	County

NOT selection:
records with
Landuse NOT Urban

ID	Area	Landuse	Municip
1	10.5	Urban	City
2	330.3	Farm	County
3	2.4	Suburban	Township
4	96.0	Suburban	County
5	22.1	Urban	City
6	30.2	Farm	Township
7	4.4	Urban	County

Figure 8-13: OR and NOT compound selections.

ations may be used to perform very complex selections.

Figure 8-14 shows the results of a complex query, combining AND, OR, and NOT operations. Here, the square brackets choose rows with a Landuse value equal to Urban, and Mill Rate values equal to B. Row 5 is the only record satisfying these criteria. The curly brackets select those rows that are not in a City, and with a Density greater than 200. This selects Row 3, and the final selected set includes both rows, by the OR operation. Selection operations may get quite complicated, and long, complex selection "sentences" may be saved, that is, the syntax copied in a text file or other repository, and applied when needed.

While database queries are typically applied to tables, we must remember that in a GIS, the tables are usually connected in some way to geographic features. Selections of table elements imply the selection of asso-

Complex selection:
records with [(Landuse = Urban) AND (Mill Rate = B)] OR
{NOT(Municip = City) AND (Density > 200)}

ID	Area	Landuse	Municip	Density	Mill Rate
1	10.5	Urban	City	1,112.2	A
2	330.3	Farm	County	1.9	C
3	2.4	Suburban	Township	237.5	C
4	96.0	Suburban	County	98.1	A
5	22.1	Urban	City	916.2	B
6	30.2	Farm	Township	3.7	A
7	4.4	Urban	County	153.8	D

Figure 8-14: An example of a complex selection, combining various selection operators.

ciated geographic elements. It is always a good idea to verify that the selection works as expected. Verification is often easiest by viewing the selection results, either on the table, the geography, or both. Figure 8-15 illustrates the results from three separate selection criteria: a) that county population be greater than 50 persons per square mile, b) that the median age be less than 40 years, and c) that housing vacancy rates be greater than 10%. The rightmost panel shows counties returned from a query specifying that criteria a and b and c all be met. The accuracy of the query may be quickly verified by inspecting maps of the component and final selections, and such an inspection should be conducted whenever possible, but especially when first learning or working with a query system.

Figure 8-16 demonstrates that queries are not generally distributive. For example, if OP1 and OP2 are operations, such as AND or NOT, then,

OP1 (ConditionA OP2 ConditionB) (8.1)

is not always the same as

(OP1 ConditionA)OP2(OP1 ConditionB) (8.2)

For example,

NOT [(Landuse = Urban) AND
(Municipality = County)] (8.3)

does not yield the same set of records as the expression

[NOT (Landuse = Urban)] AND
[NOT (Municipality = County)] (8.4)

Parentheses or other delimiters should be used to ensure unambiguous queries.

Relational databases may support a *structured query language* known as SQL (pronounced both "sequel" and "ess kyou el"). SQL was initially developed by the International Business Machines Corporation, but is supported by a number of software vendors. SQL is a nonprocedural query language in that the specification of queries does not depend on the structure of the data. The language can be powerful, general, and transferable across systems, and so has become widely adopted.

SQL provides the capability to both define and manipulate data. Data types and tables containing variables of a given type may be specified. Standard operations are

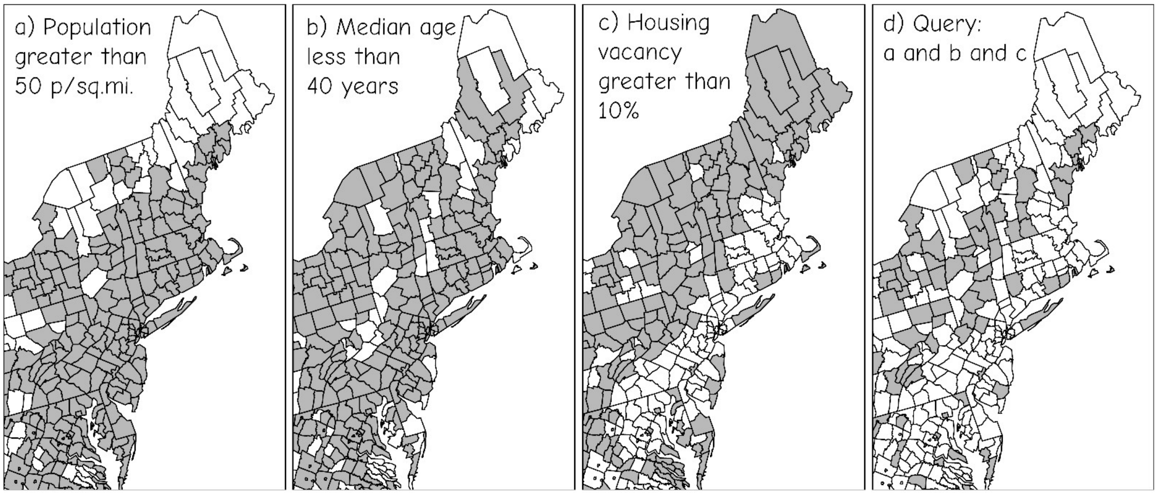


Figure 8-15: Component and composite selection criteria, applied to counties in the northeastern United States. A visual check of the composite against subcomponents is often helpful, especially when learning.

NOT [(Landuse = Urban) AND (Municip = County)]

ID	Area	Landuse	Municip
1	10.5	Urban	City
2	330.3	Farm	County
3	2.4	Suburban	Township
4	96.0	Suburban	County
5	22.1	Urban	City
6	30.2	Farm	Township
7	4.4	Urban	County

[NOT (Landuse = Urban)] AND [NOT (Municip = County)]

ID	Area	Landuse	Municip
1	10.5	Urban	City
2	330.3	Farm	County
3	2.4	Suburban	Township
4	96.0	Suburban	County
5	22.1	Urban	City
6	30.2	Farm	Township
7	4.4	Urban	County

Figure 8-16: Selection operations may not be distributed, and the order of application is very important. When the NOT operation is applied after the AND (left) a different set of records is selected than when the NOT operation is applied before the AND (right side). Order of operation is important, and ambiguity should be removed by using parentheses or other delimiters.

used to manipulate data, for example, to select, delete, insert, and update records in a database. Long or complicated queries may be saved in text files, or as scripts, that may be debugged, modified, or used later. These scripts may be quite long, and may be fairly referred to as programs, given their complexity and capabilities. Utilities can be provided to help write, test, and automate these scripts.

Because SQL as initially defined has limitations for spatial data processing, many spatial operations are not easily represented in SQL. Many more selections may be specified only with complex queries, so various SQL extensions appropriate for spatial data have been developed.

Joining Tables

Relational databases are so powerful in part because we can structure out data in ways that reduce duplication, are easier to maintain, and are flexible; much of this flexibility is because we can join tables. A join, also known as a relate, uses columns in one table to match rows based on columns in other tables. Joins were illustrated in Figure 8-7 and Figure 8-9, part d, but additional examples may help you successfully create and apply joins.

Joins are based on joint items, or join fields. In their simplest form, a single column in one table is matched to a column in another table, and a new table "created" by combining rows for matched values. Figure 8-17 shows a simple join between two tables. The simplest joins use a single column in each table as a matching or "join" item. Here, Code_A and Code_B, respectively, are used. If we call Table A the "target table," and Table B the "source table," then our join consists of matching the values for a row in the source table to the target table, and "copying" the values from the matched rows to the output table. The values aren't truly copied, but rather associated and displayed together, to save space and speed operation.

Primary Keys and Joins

We've noted earlier that keys are "key" to relational databases, and as such they have certain special characteristics, which we'll describe here. There are several kinds of keys, but the most important is the *primary key,* a chosen item or items that

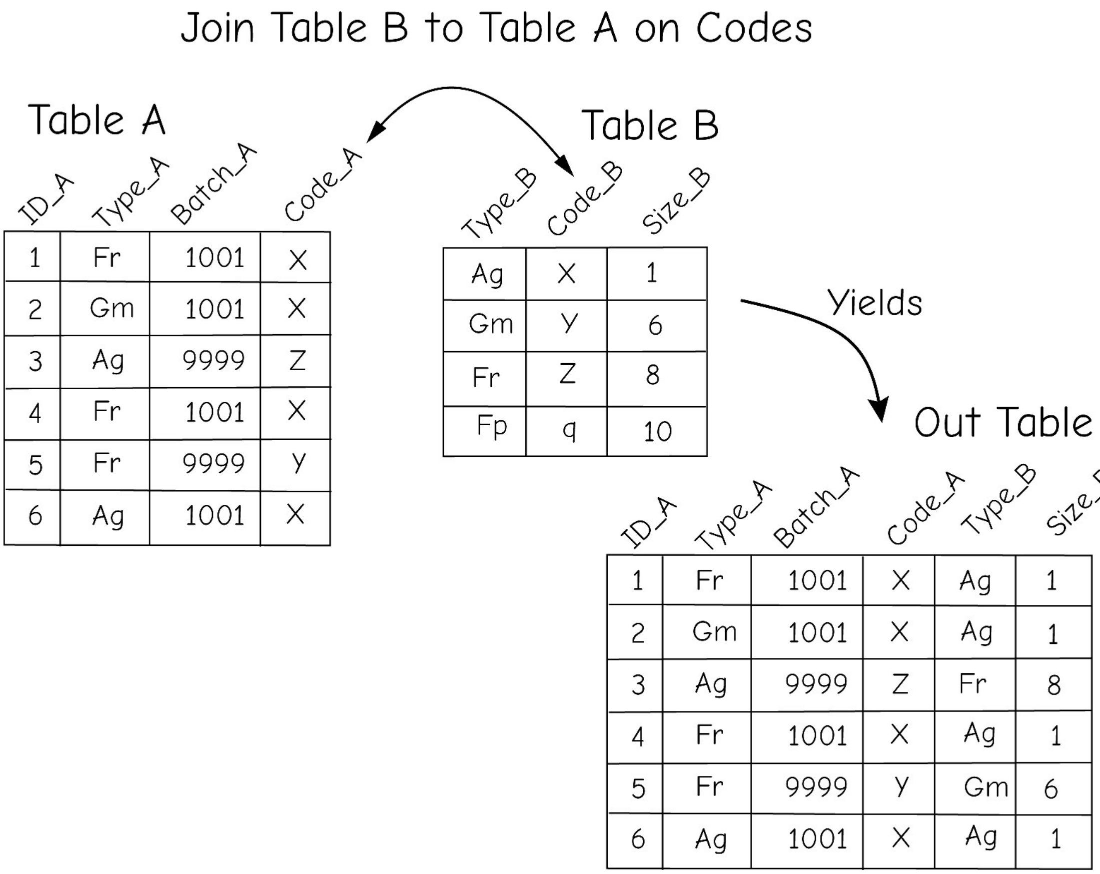


Table A

ID_A	Type_A	Batch_A	Code_A
1	Fr	1001	X
2	Gm	1001	X
3	Ag	9999	Z
4	Fr	1001	X
5	Fr	9999	Y
6	Ag	1001	X

Table B

Type_B	Code_B	Size_B
Ag	X	1
Gm	Y	6
Fr	Z	8
Fp	q	10

Out Table

ID_A	Type_A	Batch_A	Code_A	Type_B	Size_B
1	Fr	1001	X	Ag	1
2	Gm	1001	X	Ag	1
3	Ag	9999	Z	Fr	8
4	Fr	1001	X	Ag	1
5	Fr	9999	Y	Gm	6
6	Ag	1001	X	Ag	1

Figure 8-17: This figure illustrates a simple join. Table B is joined to Table A, matching the Code_B values to corresponding Code_A values, to create Out Table.

uniquely identify each row in a table. We often create and generate unique numbers as keys, for example, a social security number is unique to a person; a parts ID number is often uniquely assigned by a manufacturer; SKU numbers are unique to items in a store. ID_A can serve as a primary key for Table A in Figure 8-17 because the values uniquely identify the rows. Table B in Figure 8-17 has three columns that could serve as primary keys, Type_B, Code_B, or Size_B. These are all *candidate keys*, because each could serve as a primary key. Typically one of the candidate key columns is used as the primary key.

Figure 8-18 illustrates the join concept. When joining items, we match the corresponding rows by key values. In Step 1, Code_B, X values in Table B are matched to the Code_A, X values in Table A. The first row in Out Table is a composite of row 1 from Table A (ID_A=1, Type_A=Fr, Batch_A=1001) and row 1 from Table B (Type_B=Ag, Size_B=1). A similar matching is performed for each X in Table B, and row written in the Out Table.

Step 2 in Figure 8-18 shows a similar match for the join variable with a value of Y. Corresponding rows are matched from Table B to Table A, and written to Out Table. There is only one matching row in this case. Step 3 shows a match of the join variable Z value. Again, there is a cross-table matching, creating new rows in the Out Table. Since there are no q values in Table A, in this particular version of a join, there are no rows written in the output table.

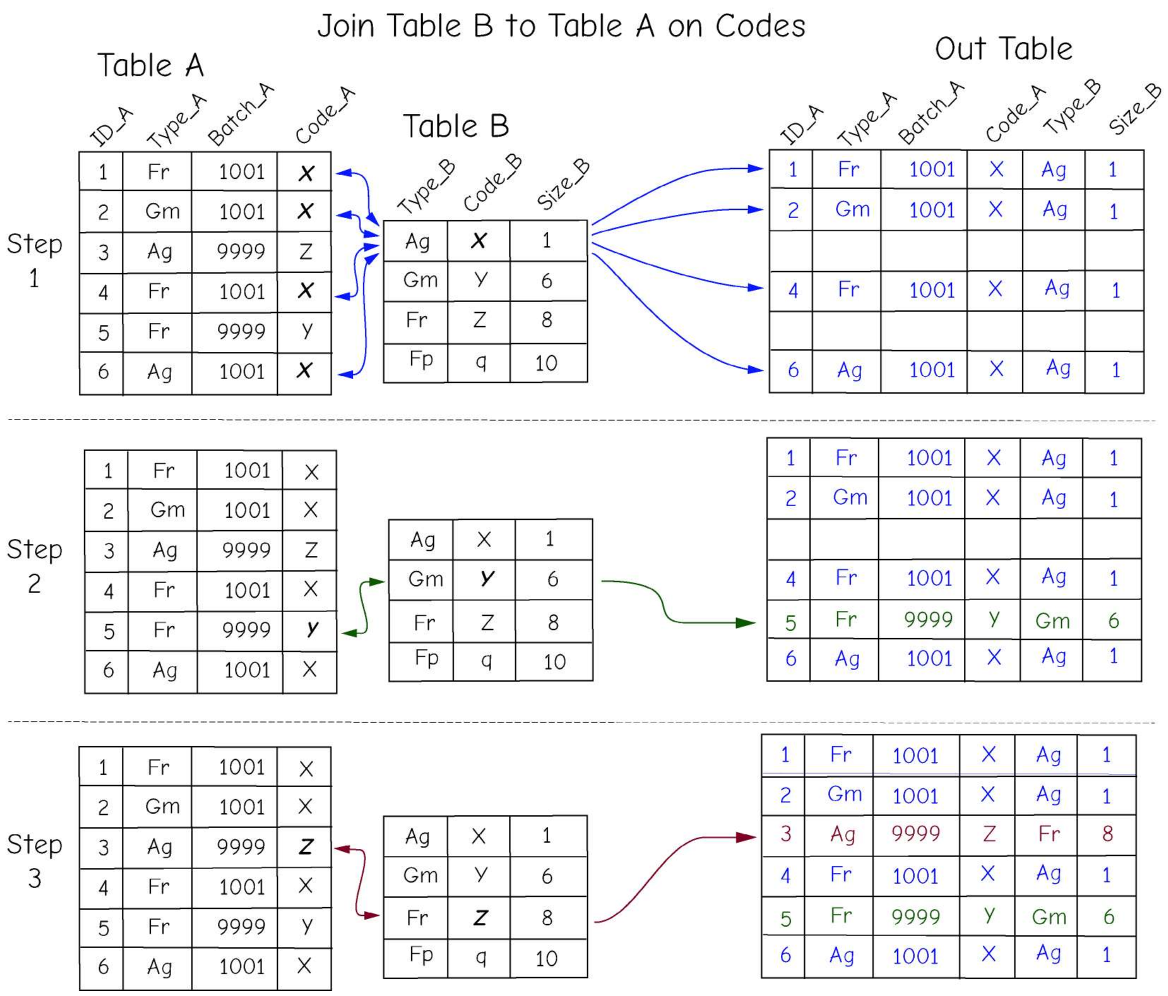


Figure 8-18: A conceptualization of how rows are combined in a join process. Join items are sequentially matched using Code_B from the "source" table (Table B) to corresponding values for Code_A in the "target" table (Table A).

Note that this example is for illustration, but database systems usually don't follow a sequential process as shown, nor write a new table. Sorts and indexing are often used to speed matches, and matched data are shown in what appears as a single table, but simply displayed from their original tables; no new table is written unless expressly requested.

Also note that non-matching elements are discarded in this example, but "missing" join values may also be retained. We may specify a join that saves some or all of the non-matching elements, and we should be aware of these different join variants.

The primary key, or an item that could serve as a primary key, is usually used as a join item in the "source" table of a join operation. This is illustrated in the join in Figure 8-17 and Figure 8-18. Code_B in the source Table B uniquely identifies each row in that table, and is used to link to Table A via Code_A. As described later, when items that are not valid as source table primary keys are used in joins, you often get ambiguous or erroneous joins.

Figure 8-17 and Figure 8-18 illustrate the most common type of join, known as an *inner join*, where unmatched rows are discarded. There are other kinds of joins. For example, an *outer join* saves the information for non-matching rows, placing blank or null values for missing items (Figure 8-19). There are both left- and right-outer joins, depending on whether the target or source nonmatches are in the target or source tables. There is also a *natural join*, in which equally named columns aren't copied, or cross joins, in which all rows in the one table are combined with all rows of another table, for example, a cross join of Tables A and B in Figure 8-17 would result in a table with 24 rows (6 rows for A times 4 rows for B).

Mastering the differences between these types of joins is perhaps a bit advanced for an introductory GIS course, but software may set any of these different types of joins as a default method, and they may not be explicitly identified by name. These and other different types of joins are covered in depth in most introductory database books, but can be confusing to distinguish and apply without some practice. I introduce them here to:

- warn you of the differences between different types of joins, and to emphasize that different types of joins will usually produce different results, even when applied to the same data, and

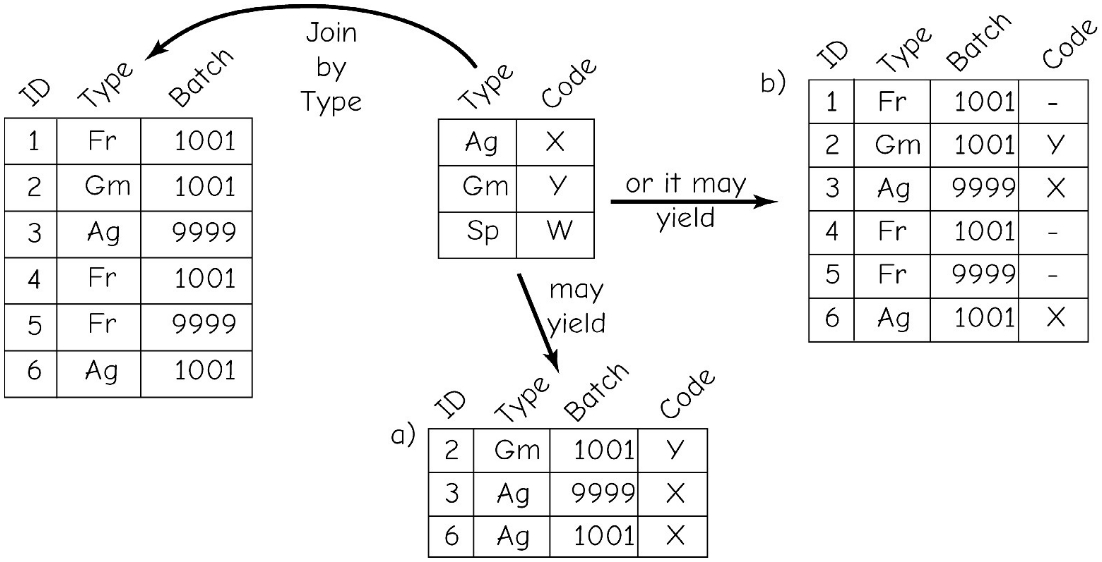


Figure 8-19: Examples of inner (a) and outer (b) joins. Note that an inner join only saves matching rows, while an outer join saves values for both matching and unmatching rows.

- to stress that there are standardized names for different types of joins, although not all GIS softwares use them. You should verify how joins work when first using new software, by comparing source and output tables.

Figure 8-19 illustrates a difference between inner and outer joins. The center table is joined to the leftmost table on the item Type. A resulting inner join is shown in a. Note that only rows 2, 3, and 6 from the "target" table on the left, with values Ag and Gm for Type (our key) are recorded in the output table in Figure 8-19a, because those are the only Type values found in both tables. Information in rows 1, 4, and 5 is not retained in the output table.

Figure 8-19b shows an *outer join*, in which unmatched source table rows are retained. Null or empty values are placed for the non-matching attributes of the source table, as shown by the dashes in the Code item for rows 1, 4, and 5 of output table b.

You may have deduced by now that the join items are crucial when joining tables. If the join items are not correctly created, then the joins will likely produce unintended results. We must pay attention to the join items across tables, particularly how many values match for our joining items across the source and target tables.

Foreign Keys

A *foreign key* is an item in a target table that may be used to unambiguously link to rows in another table. Most of our examples have used foreign keys, and it is helpful to identify them explicitly in tables so that we may maintain *referential integrity*, that is, to make sure we correctly join among tables.

We usually use a primary key or candidate key in the source table to link to a foreign key for another table (Figure 8-20). In this example, Hunter ID is the primary key of the Hunter Table. Tag ID is the primary key for the Bag Record Table. Hunter ID is a foreign key in the Bag Record Table and serves as a link to Hunter Table. This ensures that the specific harvest can be traced to the hunter, e.g. a Turkey, Bambi, and Rudolph can be unambiguously traced back to Mark Trail.

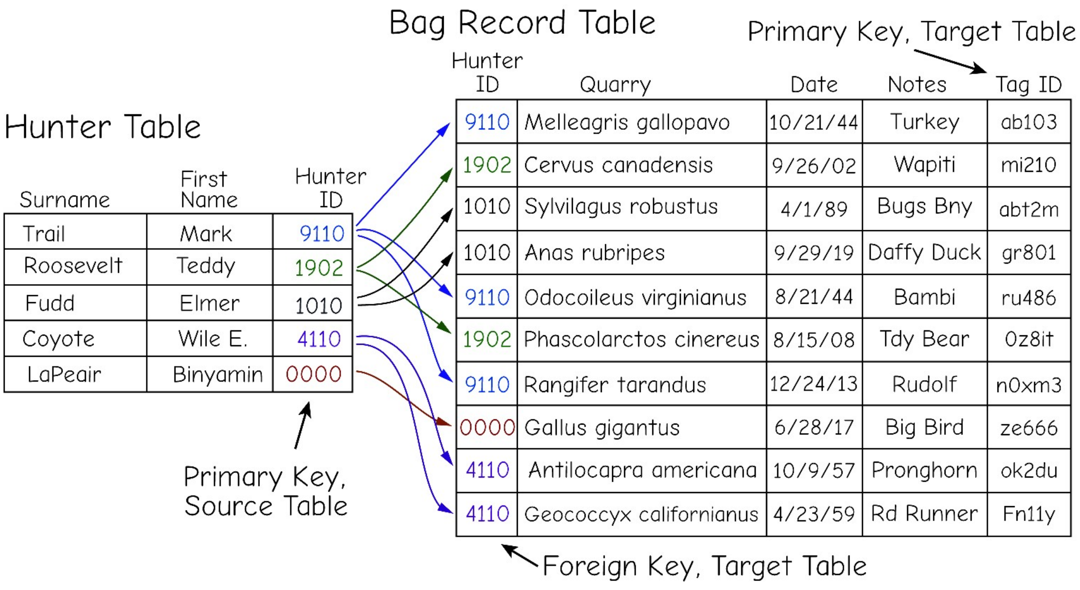


Figure 8-20: An example foreign key in a target table, used to link to a primary key in a source table. Here the Hunter ID connects the Hunter Table to the Bag Record Table. Note the Bag Record Table has a primary key, Tag ID, as well as a foreign key, Hunter ID. With apologies to Mel Blanc.

In another common example, a zip code or other postal code is often used as a primary key for a geography table, with each row in the table associated with a specific polygon in a data layer. A zip code may be added as a foreign key to another table that contains economic characteristics for each polygon. We might also have yet another table with historical population data for each zip code. Zip codes in these tables may serve as foreign keys in joins to the polygon table.

Most joins should involve a one-to-one or a one-to-many relationship between the source and the target join items. This is why we usually use a primary key in our source table as the column for a join, or we use an item that could serve as a primary key. This avoids a many-to-one relationship between source and target columns, which often results in problems. The following paragraphs illustrate these problems.

A one-to-one relationship means that there may be one and only one instance in a join item of a target table that matches one and only one instance of a join item in a source table. The left side of Figure 8-21 illustrates a one-to-one match for the items Id1 and Id2. Each value of Id2 matches only one value of Id1. Note that not all values of Id1 have a match in Id2.

Tables may also be unambiguously joined if there is a one-to-many relationship between the source join item and the target table join item. The join on the right side of Figure 8-21 shows a one-to-many relationship between the source item Id4, and the target item Id3. Note there are three instances of Y in Id3, but they unambiguously match with the one value of Y in Id4.

We often run into problems when we attempt joins with items that have a many-to-one or a many-to-many relationship. These are often considered ill-matching keys, in that results from a join can be indeterminate—you can't predict the results in advance, or they may change due to spurious factors, such as pseudo random effects of row ordering. Since you're often not sure of the results you'll get, many-to-one or many-to-many relationships from the source to target keys are rarely a good idea. We usually require the source item in a join to be a primary key or candidate key, capable of serving as a primary key: a column or set of columns that uniquely identifies the rows of the source table, so that there is a one-to-one or one-to-many relationship in a join.

Figure 8-22 shows an example of an ill-matching join item. The item Type in Source Table is not a key, and this results in a many-to-many join. There are two rows in the Source Table with a Type value of Fr. Both rows may fairly match the Fr key values found in the Target Table, resulting in an ambiguous assignment for the values of

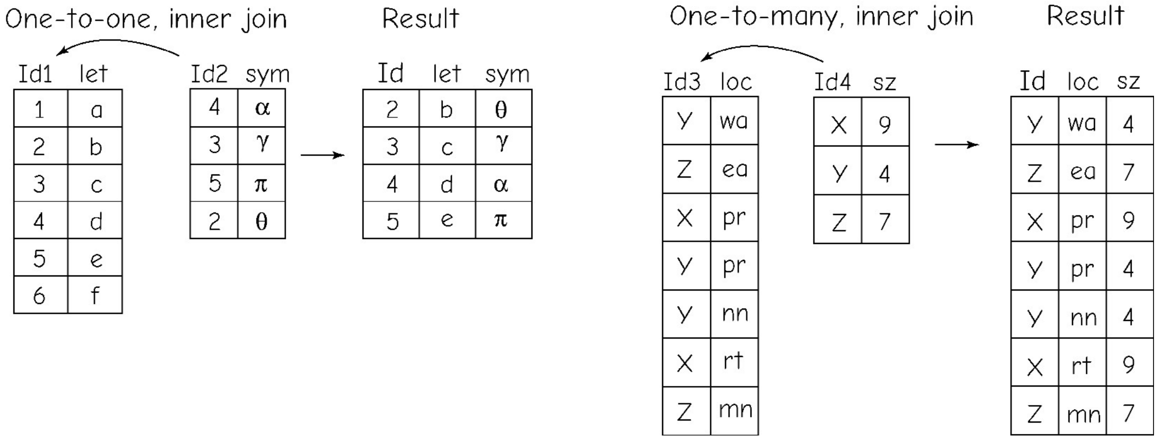
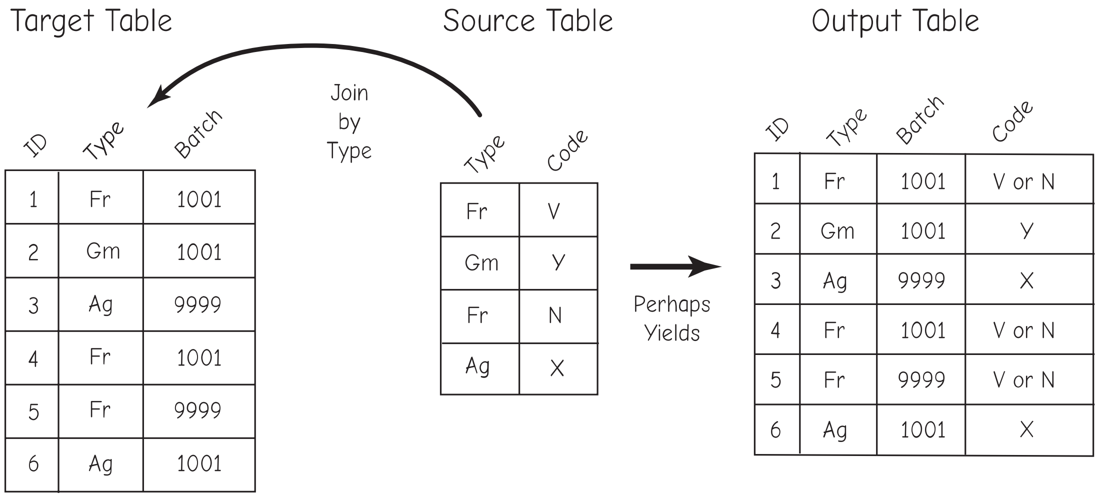


Figure 8-21: An example of a one-to-one and one-to-many relationships between tables.



Target Table

ID	Type	Batch
1	Fr	1001
2	Gm	1001
3	Ag	9999
4	Fr	1001
5	Fr	9999
6	Ag	1001

Source Table

Type	Code
Fr	V
Gm	Y
Fr	N
Ag	X

Output Table

ID	Type	Batch	Code
1	Fr	1001	V or N
2	Gm	1001	Y
3	Ag	9999	X
4	Fr	1001	V or N
5	Fr	9999	V or N
6	Ag	1001	X

Figure 8-22: An example of a many-to-many table join.

Code for those rows. Both V and N are equally supported, hence our results are uncertain, as shown in the Output Table. Such uncertainty is rarely a good thing in table manipulations or analyses. We would have the same ambiguity if there were only one value of Fr in the Target Table, creating a many to one relationship from Source to Target. Many-to-one or many-to-many joins should be avoided, except in a constrained set of circumstances.

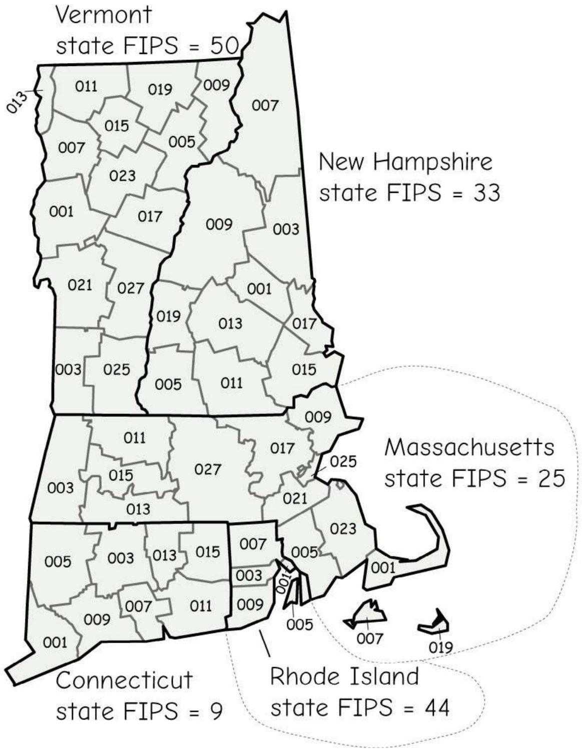
We typically are very careful about how we assign and when we alter the values of primary keys, in part because of their importance in joins. We must ensure that we don't duplicate key values within a column, and we have to be careful in how we reassign values to the primary key through calculations or other modifications. Because null values match nothing, or everything, we also don't allow them to occur in primary keys. Many database systems have checks to avoid these errors. For example, many online databases use an email address as a primary identifier, and they prevent you from registering twice under the same email address. Many of the errors in databases result from corruption of a primary key.

Concatenated Keys

Most examples show keys consisting of a single column. These are most common because they are easiest to envision, manage, and use. This is the idea behind unique identifiers in many databases, for example, an invoice number for a business, unique part ID numbers for a warehouse, or a museum accession number. These unique IDs allow unique items to be simply identified in a table.

While single-column identifiers are most common, we frequently use multiple columns as keys. These are often used when we have large, multi-table databases that we wish to combine in several different ways. Data from the United States Census, or the U.S. SSURGO database discussed in the previous chapters, use multiple tables, many with two or more columns used in combination as a key.

When multiple columns are used together as a key, it is called a *concatenated key*. Concatenated keys are typically formed by two columns, and rarely more than three columns.



STATE	COUNTY	sFIPS	cFIPS	POP (k)
Connecticut	Fairfield	9	1	882.5
Connecticut	Hartford	9	3	857.2
....	....	9	....	....
Connecticut	Windham	9	15	109.1
Massachusetts	Barnstable	25	1	222.2
Massachusetts	Berkshire	25	3	134.9
....	....	25	....	....
Massachusetts	Worcester	25	27	750.9
New Hampshire	Caroll	33	1	56.3
New Hampshire	Cheshire	33	3	43.6
....	....	33	....	....
New Hampshire	Sullivan	33	19	40.5
Rhode Island	Bristol	44	1	50.6
Rhode Island	Kent	44	3	167.1
....	....	44	....	....
Rhode Island	Washington	44	9	123.5
Vermont	Addison	50	1	35.9
Vermont	Bennington	50	3	36.9
....	....	50	....	....
Vermont	Windsor	50	27	57.4

Figure 8-23: A concatenated key, here the combined state FIPS (sFIPS) and county FIPS codes (cFIPS).

Figure 8-23 illustrates a concatenated key, here used to uniquely identify U.S. counties. Each U.S. state is assigned a unique Federal Information Processing Standard (FIPS) code, and every county within a state assigned a unique code, but unique only within the state. This allows new codes to be assigned at the state level, for example, if a new state is added, or new codes to be assigned within a state, as when counties were split to form new counties. This also allows quick selection of data by state. County FIPS codes (cFIPS) are assigned sequentially, as odd numbers within these states, from 1 up to the last county within the state. cFIPS alone can not be used as a key, for example, cFIPS = 1 specifies both Fairfield County, Connecticut, and Barnstable County, Massachusetts. A concatenated key using both state and county FIPS number is needed to uniquely distinguish counties across multiple states.

Multi-table Joins

We may have more than one potential key in a table (remember, the key can consist of one or more columns), but we usually design tables with a main key. We may also join many tables to a single table, often using different target items for each join.

Figure 8-24 shows an example of a multi-table join with distinct keys. School Table may be considered the "foundational" table, and the two tables named County and District are joined to School Table to create an Output Table.

Note that these two joins are based on different items. County Table is joined to School Table based on values in the columns labeled Cty, while the District Table is joined to the School Table based on the values found in the columns labeled DistID. Our output table is shown here without the "copies" of the columns (e.g., Cty and DistID each appear only once in the Output Table), although the "copies" are often displayed.

Note that the joins are one-to-many in both cases; one and only one value in the source columns may match many row values in the target columns. Also note that each of the join items in the source tables are keys—Cty in the County Table uniquely identifies each county, and DistID in the District Table uniquely identifies each district. As noted before, source items in joins are often keys, uniquely identifying the rows in the source table.

County Table

Name	Cty
Washoe	1
Polk	2
Placer	3
Le Seur	4

School Table

Sch#	Name	Cty	Type	DistID
11	Centenial	3	Prm	624
12	Lincoln Union	4	Hgh	624
11	Central Lakes	4	Mdl	113
13	Sierra Gardens	1	Prm	113
10	Richland	2	Prm	95
12	Nevada Union	2	Hgh	95

District Table

DistID	Superintendent
95	M. Skinner
113	D. McGee
624	R. Figgins

Output Table

Sch#	Name	Cty	Type	DistID	Name	Superintendent
11	Centenial	3	Prm	624	Placer	R. Figgins
12	Lincoln Union	4	Hgh	624	Le Seur	R. Figgins
11	Central Lakes	4	Mdl	113	Le Seur	D. McGee
13	Sierra Gardens	1	Prm	113	Washoe	D. McGee
10	Richland	2	Prm	95	Polk	M. Skinner
12	Nevada Union	2	Hgh	95	Polk	M. Skinner

Figure 8-24: An example of a multi-table join, based on different keys. Here, County and District tables are joined to the School Table, to create an Output Table.

Normal Forms in Relational Databases

Keys and Functional Dependencies

The previous sections should point out the need to carefully structure our tables in a relational database, and that keys in tables are especially important. Poorly designed tables can suffer from serious problems in performance, consistency, redundancy, and maintenance. Data that are stored in large tables may be redundant or with wasted space, and long searches may be needed to select a small set of records. Updates on large tables may be slow, and the deletion of a record may result in the unintended deletion of valuable data from the database. Smaller, carefully constructed tables are usually more useful.

Consider the data in Figure 8-25, in which building records are stored in a single table. Attributes include Parcel-ID, Alderman, Tship-ID, Tship_name, Thall-add, Own-ID, Own_name, and Own_add. Some information is stored redundantly, for example, changing the Alderman for Tship-ID 12 would require changing many rows; identifying all parcels with Yamane as an owner would require a search of all records for several columns in the table. This storage redundancy is costly both because it takes up disk space and because each extra record adds to the search and access times. A second problem comes with changes in the data. For example, if Devlin, Yamane, and Prestovic sell the parcel they jointly own (first data row), deleting the parcel record for Devlin would purge the database of her address and tax payment history. If these data on Devlin were required later, they would have to be reentered from an external source.

We may place relational databases in *normal forms* to avoid many of these problems. Data are structured in sequentially higher normal forms to improve correctness, consistency, simplicity, nonredundancy, and stability. There are several levels in the hierarchy of normal forms, but the first three levels, known as the first through third normal forms, are most common. Data are usually structured sequentially, that is, first all tables are converted to first normal forms, then converted to second and then third normal forms as needed. Prior to describing

Land Records table, unnormalized form

parcel-ID	Alderman	Tship-ID	Tship_name	Thall-add	Own-ID	Own_name	Own_add	Own-ID	Own_name	Own_add	Own-ID	Own_name	Own_add
2303	Johnson	12	Birch	15W	122	Devlin	123_pine	337	Yamane	72_lotus	890	Prestovic	12_clayton
618	DeSilva	14	Grant	35E	457	Suarez	453_highland	890	Prestovic	12_clayton	231	Sherman	64_richmond
9473	Johnson	12	Birch	15W	337	Yamane	72_lotus	-	-	-	-	-	-

Figure 8-25: Land records data in unnormalized form. The table is shown in two parts because it is too wide to fit across the page.

normal forms we must introduce some terminology and properties of relational tables.

As noted earlier, relational tables use keys to index data. There are different kinds of keys. A *super key* is one or more attributes that may be used to uniquely identify every record (row) for a table. A subset of attributes of a super key may also be a super key, and is called a *candidate key*. The *primary key* for indexing a table is chosen from the set of candidate keys. There may be many potential primary keys for a given table; however, it is usual to use only one primary key per table. The Parcel-ID is a primary key for the table in Figure 8-25, because it uniquely identifies each row in the table.

Functional dependency is another important concept. Attributes are functionally dependent if at a given point in time, each value of the dependent attribute is determined by a value of another attribute.

Figure 8-26 illustrates the concept of functional dependency. The table contains a parts list, with ID as the primary key, and a part Name, CNum, Ctype, Thread, and Angle attributes. The ID is unique for each row, and so by definition, all other items are functionally dependent on ID. If we know the value of ID is 1, then we know the part Name is Tec. We denote this as shown,

ID -> Name

We also see that Name is functionally dependent on CNum. If we know a value for CNum, say, 2, we know the value of Name will equal Ext. We see that the converse is also true here, CNum is also functionally dependent on Name. Note that this is not always true, as shown for Ctype and Thread.

CType -> Thread is true,

but

Thread -> CType is not true

Why? Because for the value of Thread equal to 14, CType may be either E or Er, violating our definition of functional dependence.

In our example in Figure 8-25, we may know that Own_add is functionally dependent on Own_name. In other words, each owner can only have one resident address, e.g., we may not allow the entry of a second resident address. Therefore, for a given Own_name, for example, Prestovic, the Own_add is determined. In a similar way, there is only one Township name,

ID	Name	CNum	CType	Thread	Angle
1	Tec	3	M	12	45
2	Cap	1	E	14	20
3	Ext	2	M	12	22
4	Cap	1	M	12	18
5	Tec	3	E	14	20
6	Cap	1	E	14	22
7	Ext	2	Er	14	45

Functional Dependencies:

ID -> Name, CNum, Ctype, Thread, Angle

CNum -> Name (or Name -> CNum)

CType -> Thread

Figure 8-26: Example of functional dependencies.

Tship_name, for each Town Hall address, Thall-add, or

Own_name -> Own_add

Tship_name -> Thall-add

Remember, these indicate that Own_add is functionally dependent on Own_name, and Thall-add is functionally dependent on Tship_name. We must always bear in mind that this functional dependency is something we enforce. Unless we place safeguards during data entry and manipulation, we may change data so that we "break" the functional dependency, for example, by adding a second owner address for an owner name.

Functional dependencies are transitive, so if A -> B, and B -> C, then A -> C. This notation means that if B is functionally dependent on A, and C is functionally dependent on B, then C is functionally dependent on A.

While relational database designs are flexible, the use of keys and functional dependencies places restrictions on relational tables:

- There cannot be repeated records, that is, there can be no two or more rows where all attributes are equal.
- There must be a primary key in a table. This key allows each record to be uniquely identified.
- No member of a column that forms part of the primary key can have a null value. This would allow multiple records which could not be uniquely identified by the primary key.

The First and Second Normal Forms

We begin creating tables in normal forms by first gathering all our data, often in a single table. Normal forms typically result in many compact, linked tables, so it is quite common to split tables as the database is *normalized*, or placed in normal forms. After normalization, the tables have an indexing system that speeds searches and isolates values for updating.

Tables with repeat groupings, as in the table at the top of Figure 8-27, are *unnormalized*. A repeating group exists in a relational table when an attribute is allowed to have more than one value represented within a row. Owner-ID repeats itself for dwellings with multiple owners.

A table is in first normal form when there are no repeat columns. The Land Records table at the bottom of Figure 8-27 has been *normalized* by placing each owner into a separate row. This is a table in the first normal form (1NF) because each column appears only once in the table definition. A 1NF is the most basic level of table normalization. However, the 1NF table structure still suffers from excessive storage redundancy, inefficient searches, and potential loss of data on updating. First normal forms have an advantage over unnormalized tables because queries are easier to code and implement. Tables in 1NF are usually converted to higher-order normal forms, usually to at least third normal form, 3NF, but it is useful to understand second normal forms before describing 3NF tables.

A table is in second normal form (2NF) if it is in first normal form and every non-key attribute is functionally dependent only on the primary key, or on transitive functional dependencies of the primary key. Remember that functional dependency means that knowing the value for one attribute of a record automatically specifies the value for the functionally dependent attribute. The non-key attributes may be directly depen-

Land Records table, unnormalized form

parcel-ID	Alderman	Tship-ID	Tship_name	Thall-add	Own-ID	Own_name	Own_add	Own-ID	Own_name	Own_add	Own-ID	Own_name	Own_add
2303	Johnson	12	Birch	15W	122	Devlin	123_pine	337	Yamane	72_lotus	890	Prestovic	12_clayton
618	DeSilva	14	Grant	35E	457	Suarez	453_highlar	890	Prestovic	12_clayton	231	Sherman	64_richmond
9473	Johnson	12	Birch	15W	337	Yamane	72_lotus	-	-	-	-	-	-

Land Records table, first normal form (1NF)

parcel-ID	Alderman	Tship-ID	Tship_name	Thall-add	Own-ID	Own_name	Own_add
2303	Johnson	12	Birch	15W	122	Devlin	123_pine
2303	Johnson	12	Birch	15W	337	Yamane	72_lotus
2303	Johnson	12	Birch	15W	890	Prestovic	12_clayton
618	DeSilva	14	Grant	35E	457	Suarez	453_highland
618	DeSilva	14	Grant	35E	890	Prestovic	12_clayton
618	DeSilva	14	Grant	35E	231	Sherman	64_richmond
9473	Johnson	12	Birch	15W	337	Yamane	72_lotus

Figure 8-27: Relational tables in unnormalized (top) and first normal forms (bottom).

dent on the primary key through some functional dependency, or they may be dependent through a transitive dependency. The Land Records table in 1NF at the bottom of Figure 8-27 has only one possible primary key, the composite of Parcel-ID and Own-ID. No other combination uniquely identifies each row. However, this table is not in second normal form because it has non-key attributes that are not functionally dependent only on the primary key attributes. For example, Tship_name and Thall_add are functionally dependent on Tship-ID.

The Land Records table at the bottom of Figure 8-27 is repeated at the top of Figure 8-28. This table exhibits the primary disadvantages of the first normal form. Parcel-ID, Alderman, and Tship-ID are duplicated when there are multiple owners of a parcel, causing burdensome data redundancy. Each time these records are updated, for example when a new Alderman is elected, data must be changed for each duplicate record. If a parcel changes hands and the seller does not own another parcel represented in the table, then information on the seller is lost.

Some of these disadvantages can be removed by converting the first normal form table to a group of second normal form tables. To create second normal form tables, we make every non-key attribute fully dependent on a primary key in the new tables. Note that the 1NF table will often be split into two or more tables when converting to 2NF, and each new table will have its own key. Any non-key attributes in the new tables will be dependent on the primary keys. The bottom of Figure 8-28 shows our Land Records converted to second normal form. Each of the three tables in second normal form isolates an observed functional dependency, so each table and dependency will be described in turn.

How do we systematically apply this criterion that the non-key attributes be functionally dependent only on the primary key, directly or through a transitive functional dependency? We must 1) specify the primary key, 2) identify the main functional dependencies, and 3) project the 1NF table across the key and dependency columns.

First, we must identify the primary key. In our example here, the simplest primary key is the (concatenated) key that is the combination of Parcel-ID and Owner-ID. If our primary key is a single item then the table is already in 1NF by definition, because all non-key attributes will depend on the primary key. However, if our primary key is more than one column, we may have further work to convert to 2NF, focusing on dependence on the components of the primary key.

Our second step is to identify the functional dependencies.We know that parcels occur in only one township, and that each

Land records table, first normal form (1NF)

parcel-ID	Alderman	Tship-ID	Tship_name	Thall-add	Own-ID	Own_name	Own_add
2303	Johnson	12	Birch	15W	122	Devlin	123_pine
2303	Johnson	12	Birch	15W	337	Yamane	72_lotus
2303	Johnson	12	Birch	15W	890	Prestovic	12_clayton
618	DeSilva	14	Grant	35E	457	Suarez	453_highland
618	DeSilva	14	Grant	35E	890	Prestovic	12_clayton
618	DeSilva	14	Grant	35E	231	Sherman	64_richmond
9473	Johnson	12	Birch	15W	337	Yamane	72_lotus

Given functional dependencies:

Parcel-ID → Alderman, Tship-ID
Tship-ID → Tship_name, Thall-add
Own-ID → Own_name, Own_add

Land records tables, second normal form (2NF)

Land Records Table 1

parcel-ID	Alderman	Tship-ID	Tship_name	Thall-add
2303	Johnson	12	Birch	15W
618	DeSilva	14	Grant	35E
9473	Johnson	12	Birch	15W

Land Records Table 2

Own-ID	Own_name	Own_add
122	Devlin	123_pine
337	Yamane	72_lotus
890	Prestovic	12_clayton
457	Suarez	453_highland
231	Sherman	64_richmond

Land Records Table 3

parcel-ID	Own-ID
2303	122
2303	337
2303	890
618	457
618	890
618	231
9473	337

Figure 8-28: Ownership data, converted to second normal form.

township has a unique Tship-ID, a unique Tship_name, a unique Thall_add, and one Alderman. This means that if we have identified a parcel by its Parcel-ID, the Alderman, Tship-ID, Tship_name, and Thall_add are known. We assign a unique identifier to each parcel of land, and the Alderman, Tship_name, and Thall_add are all dependent on this identifier. This means if we know the parcel identifier, we know these remaining values. This is the definition of functional dependency. We represent these functional dependencies by:

Parcel-ID -> Alderman

Parcel-ID -> Tship-ID

Parcel-ID -> Tship_name

Parcel-ID -> Thall_add

These functional dependencies are incorporated in the table named Land Records 1 in Figure 8-28.

Second, note that once Own-ID is specified, the Own_name and Own_add are determined. Each owner has a unique identifier and only one name (aliases not allowed). Also, each owner has only one permanent home address. Own_name and Own_add are functionally dependent on Own-ID. The functional dependencies are:

Own-ID -> Own_name

Own-ID -> Own_add

The Parcel-ID and Own-ID are called partial functional dependencies, because while both are dependent on the primary key, they aren't dependent on each other. If I have a unique Parcel-ID, I know additional information about some of the columns for any row in the table, but not all of the columns. If I know the Own-ID, I also know the values of a set of columns, but again, not all. When we have a concatenated key, we must identify these in our data, and they guide us in how to further split our table.

How do we get to 2NF? By projecting the 1NF table across the primary key and functional dependencies. Remember, project is just a way of saying we subset the columns, here guided by the functional dependencies. These partial functional dependencies are represented in the tables Land Records 1 and Land Records 2 in Figure 8-28.

Finally, note that we need to tie the owners to the parcels. These relationships are presented in the table Land Records 3 in Figure 8-28. Note that some parcels are jointly owned, and so there are multiple owner IDs for each parcel.

The three tables Land Records 1 through 3 satisfy the conditions of a second normal form. Second normal form eliminates some of the redundancies associated with the 1NF. Note that the redundancy in storing the information on Alderman, Tship-ID, Tship_name, and Thall_add have been significantly reduced, and the minor redundancy in Own_name has also been removed. Editing the tables becomes easier; for example, changes in Alderman entail modifying fewer records. Finally, deletion of a parcel does not have the side effect of deleting the information on the owner, Own-ID, Own_name, and Own_add.

The Third Normal Form

The 2NF still contains problems, although they are small compared to a table in 1NF. They can still suffer from transitive functional dependencies. If a transitive functional dependency exists in a table, then there is a chain of dependencies. A transitive dependency occurs in our example table named Land Records 1 (Figure 8-28). Note that Parcel-ID specifies Tship-ID, and Tship-ID specifies Tship_name, Thall_add and Alderman. In our notation of functional dependencies:

Parcel-ID - > Tship-ID

and

Tship-ID -> Tship_name, Thall_add, Alderman

This causes a problem when we delete a parcel from the database. To delete a parcel we remove the parcel from tables Land Records 1 and Land Records 3. In so doing, we might also lose the relationship among Tship-ID, Tship_name, Thall_add, and Alderman. To avoid these problems we need to convert the tables to the third normal form.

A table is in the third normal form (3NF) if and only if for every functional dependency A -> B, A *is a super key, or* B *is a member of a candidate key.* This requirement means we must identify transitive functional dependencies and remove them, typically by splitting the table that contains them. The tables Land Records 2 and Land Records 3 in Figure 8-28 are already in the 3NF, because the keys for these tables are super keys. Owner-ID uniquely identifies the rest of the row in Land Records 2, and the concatenated key of Parcel-ID and Tship-ID are the rows in Land Records 3.

However, the table Land Records 1 in Figure 8-28 is not in 3NF because the functional dependencies for table Land Records 1 are:

Parcel-ID - > Tship-ID

Tship-ID -> Tship_name, Thall_add, Alderman

Tship-ID is not a super key for the table, nor are Tship_name and Thall_add members of a primary candidate key for that table. Removing the transitive functional dependency by splitting the table will create two new tables, each of which satisfies the criteria for the 3NF. Figure 8-29 contains the tables Land Records 1a and Land Records 1b, both of which now satisfy the 3NF criteria, and preserve the information contained

Land records, third normal form

Land Records 1a

FD: Parcel-ID → Tship-ID

Parcel-ID	Tship-ID
2303	12
618	14
9473	12

Land Records 1b

FD: Tship-ID → Tship_name, Thall_add, Alderman

Tship-ID	Tship_name	Thall_add	Alderman
12	Birch	35W	Johnson
14	Grant	35E	DeSilva

Land Records 2

FD: Own-ID → Own_name, Own_add

Own-ID	Own_name	Own_add
122	Devlin	123_pine
337	Yamane	72_lotus
890	Prestovic	12_clayton
457	Suarez	453_highland
231	Sherman	64_richmond

Land Records 3

No Functional Dependencies

Parcel-ID	Own-ID
2303	122
2303	337
2303	890
618	457
618	890
618	231
9473	337

Figure 8-29: Ownership data in third normal form, with the functional dependencies (FD) noted at the top of the table.

in the 1NF table in Figure 8-27. Note that Parcel-ID is now a super key for Table 1a and Tship-ID is a super key for Table 1b, so the 3NF criteria are satisfied.

A general goal in defining a relational database structure is to have the fewest tables possible that contain the important relationships and to have all tables in at least 3NF. Normal forms higher than three have been described and provide further advantages; however, these higher forms are often more limited in their application and depend on the intended use of the database.

While relational tables in normal forms have certain useful characteristics, they may suffer from relatively long access times for specific queries. Databases may be organized around usage, or *denormalized* for the most common processes. These denormalizations typically add extra columns or permanent joins to the database structure. This may add redundancy or move a table to a lower normal form, but these disadvantages often allow significant gains in processing speed. The need to denormalize tables has diminished with improvements in computing power. However, denormalization may be required for extremely large databases, or where access speed is of primary importance.

Summary

Attribute data are an important component of spatial data in a GIS. These data may be organized in several ways, but data structures that use relational tables have become the most common method for organizing and manipulating attribute data in GIS.

Selections, or queries, are among the most common analyses conducted on attribute data. Queries mark a subset of records in a table, often as a precursor to subsequent analyses. Queries may use AND, OR, and NOT operations, among others, alone or in combination.

Keys are important in structuring relational data tables. Primary and foreign keys are defined which are used in joining tables. Primary keys may sometimes be concatenated, or formed from several columns, rather than just from one column. Entry and manipulation of key values is often constrained so that tables may function properly.

Relational tables are often placed in normal forms to improve correctness and consistency, to remove redundancy, and to ease updates. Normal forms seek to break large tables into small tables that contain simple functional dependencies. This significantly improves the maintenance and integrity of the database. Normal forms may cause some cost in speed of access, although this is a diminishing problem as computer hardware improves.

Object-relational database systems have been developed that incorporate the strong typing and domains of object-oriented models with the flexibility, logic, and ubiquity of relational data models. These evolutionary improvements to the relational approach will continue as database technologies are extended across networks of computers and the World Wide Web.

Suggested Reading

Adam, N., Gangopadhyay, A. (1997). *Database Issues in Geographic Information Systems*. Dordrecht: Kluwer Academic Publishers.

Adler, D.W. (2000). *IBM DB2 Spatial Extender - Spatial data within a RDBMS*. Proceedings 27th International Conference on Very Large Databases. Rome, Italy.

Arctur, D., Zeiler, M. (2004). *Designing Geodatabases: Case Studies in GIS Data Modeling*. Redlands: ESRI Press.

Aronoff, S. (1991). *Geographic Information Systems: A Management Perspective*. WDL Publications: Ontario.

Bhalla, N. (1991). Object-oriented data models: a perspective and comparative review. *Journal of Information Science*, 17:145–160.

Codd, E.F. (1982). Relational database: a practical foundation to productivity. *Communications of the ACM*, 25:109–117.

Date, C.J. (2004). *An Introduction to Database Systems* (8th ed.). Boston: Pearson/ Addison-Wesley.

Frank, A.U. (1988). Requirements for a database management system for a GIS. *Photogrammetric Engineering and Remote Sensing*, 54:1557–1564.

Lorie, R.A., Meier, A. (1984). Using a relational DBMS for geographical databases. *Geoprocessing*, 2:243–257.

Milne, P., Milton, S., Smith, J.L. (1993). Geographical object-oriented databases: a case study. *International Journal of Geographical Information Systems*, 7:39–55.

Obe, R.O., Hsu, L.S. (2011). *PostGIS in Action*. Greenwich: Manning Publications.

Rigaux, P., Scholl, M., Voisard, A. (2002). *Spatial Databases With Applications To GIS*. San Francisco: Morgan Kaufman.

Teorey, T.J. (1999). *Database Modeling and Design* (3rd ed.). San Francisco: Morgan Kaufmann.

Ullman, J.D., Widom, J. (2008). *A First Course in Database Systems.* New York: Prentice Hall.

Zeiler, M. (1999). *Modeling Our World: An ESRI Guide to Geodatabase Design*. Redlands: ESRI Press.

Study Questions

8.1 - What are the main components of a database management system?

8.2 - What are the primary functions of a database management system?

8.3 - Describe the difference between single and multiple user views.

8.4 - What is a one-to-one relationship between tables? A many-to-one relationship?

8.5 - Which single columns in the following table may serve as keys?

PID	Osel	Clr	NumT	SpLm
1	B	or	1	55
3	D	gr	2	55
5	A	rd	11	55
7	C	ye	23	55
9	G	az	1	65
null	X	bl	9	65

8.6 - Which single columns in the following table may serve as keys?

CID	TStmp	Osel	Clr	NumT	Xerr
1	10:12	B	rd	1	110
3	11:44	D	gr	-5	220
5	11:44	A	rd	11	220
7	16:58	C	gr	23	110
9	22:11	F	bl	0	110
Null	23:59	H	bl	-2	220

8.7 - Why have relational database structures proven so popular?

8.8 - What are the eight basic operations formally defined by E.F. Codd for the relational model?

8.9 - What is the primary reason that hybrid database models are used for spatial data?

8.10 - Does an OR condition result in more, fewer, or the same number of records than the component parts? For example, is the set from:

condition A OR condition B

the same, bigger, or smaller than the set from condition A alone, or condition B alone?

8.11 - Does an AND condition result in more, fewer, or the same number of records as the component parts? For example, is the set from:

condition A AND condition B

the same, bigger, or smaller than the set from condition A alone, or condition B alone?

8.12 - Identify the states meeting each of the following selection criteria, based on the table below:

a) Smokers < 20%

b) Smokers > 20% and illiteracy < 10

c) Not (non-federal taxes > 9)

d) Illiteracy < 7 or income > 22,000

e) Get more federal aid than paid in taxes, and non-federal taxes > 9

f) [Firearms deaths < 10 and income > 21,000] and not {smokers > 20}

FIPS	Name	Smokers (%)	Income ($/person)	Illiteracy (%)	Firearm deaths / 100,000	Non-Federal Tax Rate (%)	Fed. Aid / Fed. Taxes
01	Alabama	22,1	18,189	15	16.2	8.6	1.71
02	Alaska	21.5	22,660	9	20	6.4	1.87
12	Florida	17.5	21,557	20	11.1	7.4	1.02
13	Georgia	19.5	21,154	17	13.4	9.9	0.96
19	Iowa	18.8	19,674	7	6.7	9.7	1.11
27	Minnesota	17.6	23,198	6	6	10.2	0.69
40	Oklahoma	24.7	17,646	12	13.1	9.8	1.48
55	Wisconsin	19.9	21,271	7	8.1	10.2	0.82

8.13 - Identify the countries from the following table that meet the following criteria.

a) Per capita energy use > 4,000 and population < 20,000,000

b) Infant mortality < 7 and life expectancy > 79.0

c) Per capita energy use < 4,000 or ((population > 40 million) and (car theft < 1))

d) [Per capita energy use < 4,000 or (population > 40 million)] and (car theft < 1)

e) not (population > 40,000,000)

f) Population < 20,000,000 and not (car theft > 1.5)

Country	Population (millions)	Energy Use (bl.oil/per)	Infant Mortality (per 1000)	Life expect. (years)	Car Theft (%)
Australia	19.9	5,668	4	79.2	2.2
Britain	59.3	5,945	5	77.5	2.6
Finland	5.2	6,456	4	78.0	0.5
France	59.7	4,350	4	79.2	1.8
Japan	127.2	4,071	3	81.6	0.1
Netherlands	16.2	5,993	5	78.3	0.5
Norway	4.6	6,019	4	78.9	1.5
South Africa	45.3	3,703	52	46.5	2.4
Spain	41.1	2,945	5	78.3	0.5
U.S.A.	291.0	8,066	7	77.3	0.5

8.14 - What are normal forms in relational databases? Why are they used, and what are the advantages of putting data in higher normal forms?

8.15 - Sketch the output table resulting from an inner join shown below:

Id1	pos
Y	wa
Z	ea
A	rt
Y	pr
Y	nn
R	rt
Q	mn

Id2	tm
X	5
Y	1
A	6
Q	4
N	3
L	2

8.16 - Sketch an outer join for the table shown in the previous problem.

8.17 - Define the basic differences between first, second, and third normal forms.

8.18 - Give an example of a functional dependency.

8.19 - List the single-column functional dependencies in the following table, using the arrow notation described in this chapter.

ID	Size	Shape	Color	Age	Source
1	large	round	blue	10	A
2	medium	round	green	5	B
3	small	round	red	10	C
4	medium	knobbed	green	5	D
5	medium	knobbed	green	5	E
6	large	round	blue	10	F
7	large	round	blue	10	A

8.20 - List the single-column functional dependencies in the following table, using the arrow notation described in this chapter.

ID	Size	Shape	Color	Age	Source
5	medium	round	blue	5	A
2	large	round	green	10	B
3	small	round	green	10	C
4	large	knobbed	green	10	D
5	medium	knobbed	blue	5	E
5	medium	round	blue	10	F
7	large	round	green	10	G

9 Basic Spatial Analyses

Introduction

Spatial data analysis is the application of operations to coordinate and related attribute data, often to solve a problem. We may wish to identify high crime areas, to generate a list of road segments that need repaving, or find the best area to place wind turbines. There are hundreds of *spatial operations* or *spatial functions* used in spatial analysis, and all involve calculations with coordinates or attributes.

Spatial operations are often applied sequentially to solve a problem, the output of each spatial operation serving as the input of the next (Figure 9-1). Part of the challenge in geographic analysis is selecting appropriate spatial operations, and applying them in the appropriate order.

The table manipulations we described in Chapter 8 are included in our definition of a spatial operation. Indeed, the selection and modification of attribute data in spatial data layers are included at some time in nearly all complex spatial analyses. Many operations incorporate both the attribute and coordinate data, and the attributes must be further selected and modified in the course of a spatial analysis.

The discussion in the present chapter will expand on rather than repeat the selection operations treated in Chapter 8. This chapter describes spatial data analyses that involve sort, selection, classification, and spatial operations that are applied to both coordinate and associated attribute data.

Input, Operations, and Output

Spatial data analysis typically involves using data from one or more layers to create output. The analysis may consist of a single operation applied to a data layer, or many operations that integrate input data from many layers to create the desired output.

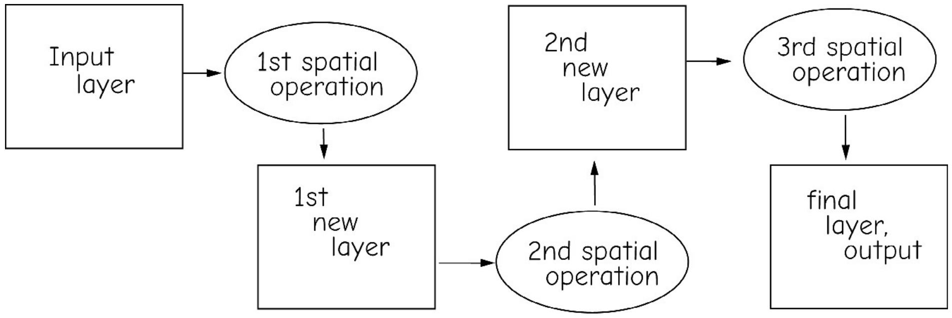


Figure 9-1: A sequence of spatial operations is often applied to obtain a desired final data layer.

There are also operations that generate several output data layers from a single input. Terrain analysis functions may take a raster grid of elevations as an input data layer and produce both slope (local steepness) and aspect (the slope direction). In this case, two outputs are generated for each input elevation data set.

Operations may also take several input layers to generate a single output layer. A layer average is an example of the use of multiple input layers to produce a single output layer, for example, annual rainfall found by summing 12 monthly rainfall raster layers. Finally, there are some spatial operations that require many input layers and generate many output layers.

The output from a spatial operation may be spatial, creating new spatial data layers, or nonspatial, producing scalar values or a table, with no explicit geometric data attached. A layer average function may simply calculate the mean cell value found in a raster data layer. The input is a spatial data layer, but the output is a single number.

Scope

Spatial data operations may be characterized by their *spatial scope*, the extent or area of the input data that are used in determining the values at output locations (Figure 9-2). Spatial operations may be characterized as local, neighborhood, or global, to reflect the extent of the input area used to determine the value at a given output location.

Local operations use only the data at one input location to determine the value at that same output location (Figure 9-2, top). Attributes or values at adjacent locations are not used in the operation.

Neighborhood operations use data from both an input location plus nearby locations to determine the output value (Figure 9-2, center). The extent and relative importance of values in the nearby region may vary, but the value at an output location is influenced by more than just the value of data found at the corresponding input location.

Global operations use data values from the entire input layer to determine each output value. The value at each location depends in part on the values at all input locations (Figure 9-2, bottom).

The set of available spatial operations depends on the data model and type of spatial data used as input. Some operations may be easily applied to raster or vector data. While the details of the specific implementation may change, the concept of the operation does not. Other operations may be possible in only one data model.

Characteristics of a data model will determine how any given operation is applied. The specific implementation of many operations, for example, multilayer addition, depends on the specific data model. A raster operation may produce a different outcome than a vector operation, even if the themes are meant to represent the same features. In a like manner, the specific set and sequence of operations in a spatial analysis will depend on the data model used and the specific operations available in the GIS software.

Spatial scope provides a good example of this influence of data models. Cells in a raster data set have uniform size and shape. A local operation applied to a raster data layer has a well-defined, repeatable area. In contrast, polygons usually vary in size and shape. A local operation for a vector polygon data set is likely to have variable size and shape from one location to another. In Figure 9-2, the local operation follows a state boundary. Therefore, the operation applies to a different size and shape for each state.

Neighborhood analyses are affected by the shape of adjacent states in a similar manner. Summary values such as populations of adjacent states may be greatly influenced by changes in neighborhood size, so great care must be taken when

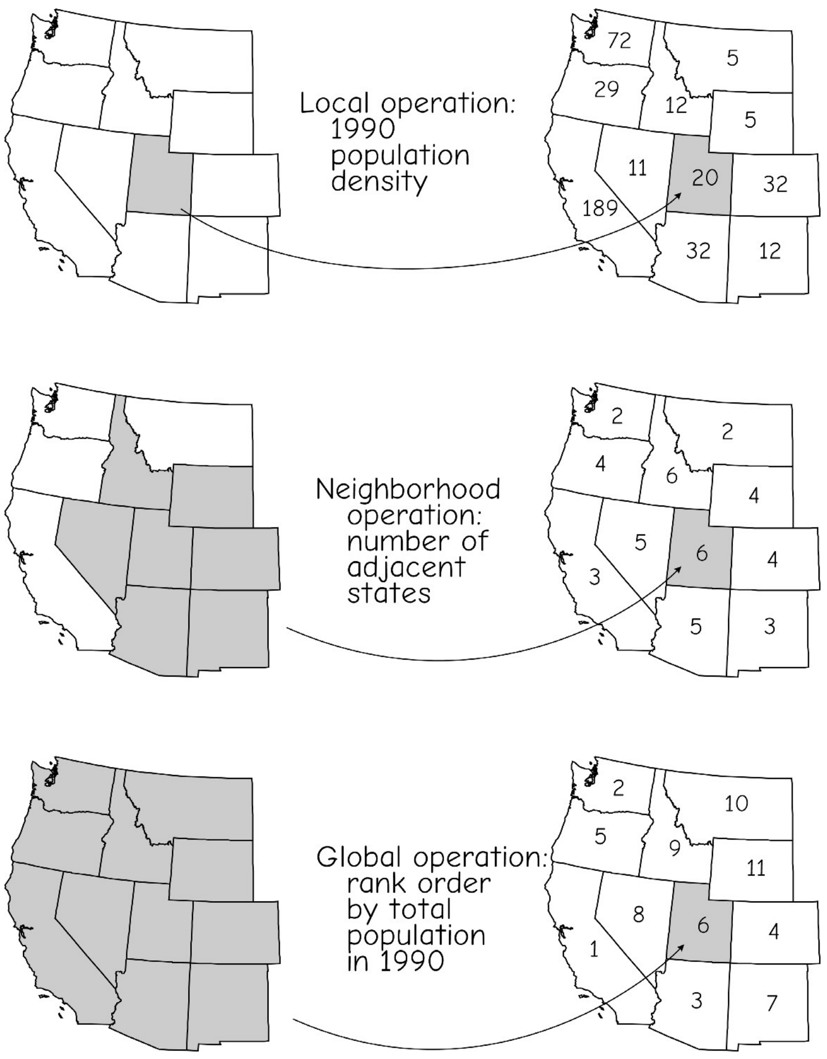


Figure 9-2: Local, neighborhood, and global operations. Specific input and output regions are shown for Utah, the shaded area on the right side of the figure. Shaded areas on the left contribute to the values shown in the shaded area on the right. Local operation output (top right) depends only on data at the corresponding input location (top left). Neighborhood operation output (middle right) depends on input from the local and surrounding areas (middle left). Global operation output (bottom right) depends on all features in the input data layer (bottom left).

interpreting the results of a spatial operation. Knowledge of the algorithm behind the operation is the best aid to interpreting the results.

While most operations might be conceptually compatible with most spatial data models, some operations are easier to apply in some models. Most neighborhood operations are quite easy to program when using raster data models, and quite difficult when using vector data models. The reverse is true for network operations, which are generally easier to apply in vector models. In many instances, it is more efficient to convert the data between data models and apply the desired operations and, if necessary, convert the results back to the original data model.

Selection and Classification

Selection operations identify features that meet one to several conditions or criteria. In these operations, attributes or geometry of features are checked against criteria, and those that satisfy the criteria are selected. These selected features may then be written to a new output data layer, or the geometry or attribute data may be manipulated in some manner.

Figure 9-3 shows an example of a selection operation that involves the attributes of a spatial data set. Two conditions are applied, and the features that satisfy both conditions are included in the selected set.This example shows the selection of those states in the "lower 48" United States that are a) entirely north of Arkansas, and b) have an area greater than 84,000 km^2. The complete set of features that will be considered is shown at the top of the figure. This set is composed of the lower 48 states, with the state of Arkansas indicated by shading. The next two maps of Figure 9-3 show those states that match the individual criteria. The second map from the top shows those states that are entirely north of Arkansas, while the third map shows all those states that are greater than 84,000 km^2. The bottom part of Figure 9-3 shows those states that satisfy both conditions. This figure illustrates two basic characteristics of selection operations. First, there is a set of features that are candidates for selection, and second, these features are selected based singly or on some combination of the geographic and attribute data.

The simplest form of selection is an *on-screen query*. A data layer is displayed, and features are selected by a human operator. The operator uses a pointing device to locate a cursor over a feature of interest and sends a command to select, often via a mouse click or keyboard entry. On-screen (or interactive) query is used to gather information about specific features, and is often used for interactive updates of attribute or spatial data. For example, it is common to set up a process such that when a feature is selected, the attribute information for the feature is displayed. These attribute data may then be edited and the changes saved.

Queries may also be specified by applying conditions solely to the aspatial components of spatial data. These selections are most often based on the attribute data tables for a layer or layers. These selection operations are applied to a set of features in a data layer or layers. The attributes for each feature are compared to a set of conditions. If the attributes match the conditions, they are selected; if the attributes fail to match the conditions, they are placed in an unselected set. In this manner, selection splits the data into either the selected set or the unselected set. The selected data are then typically acted on in some way, often saved to a separate file, deleted, or changed in some manner.

Selection operations on tables were described in general in Chapter 8. The description here expands on that information and draws attention to specific characteristics of selections applied to spatially related data. Table selections have spatial relevance because each record in a table is associated with a geographic feature. Selecting a record in a table simultaneously selects the associated spatial features: cells, points, lines, or areas. Spatial selections may be combined with table selections to identify a set of selected geographic features.

Set Algebra

Selection conditions are often formalized using *set algebra*. Set algebra uses the operations less than (<), greater than (>), equal to (=), and not equal to (< >). These selection conditions may be applied either alone or in combination to select features from a set.

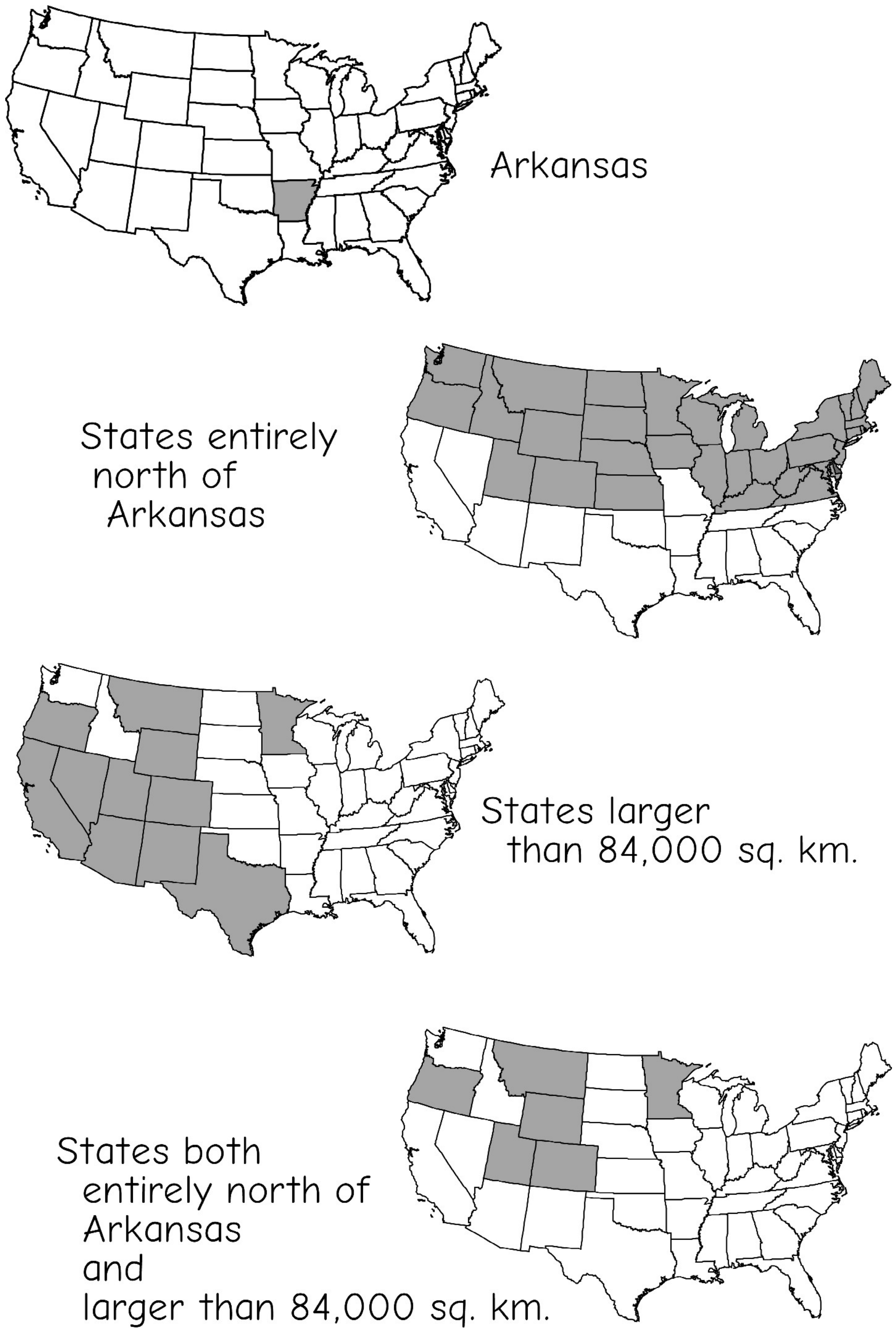


Figure 9-3: An example of a selection operation based on single or multiple conditions.

Figure 9-4 shows four set algebraic expressions and the selection results for a set of counties in the northeastern United States. The upper two selections show equal to (=) and not equal to (< >) selections. The upper left shows all counties with a value for the attribute state that equals Vermont, while the upper right shows all counties with a value for state that are not equal to New York. The lower selections in Figure 9-4 show examples of ordinal comparisons. The left figure shows all counties with a size greater than or equal to (> =) 1,000 mi^2, while the right side shows all counties with a population density less than (<) 250 persons per mi^2.

The set algebra operations greater than (>) or less than (<) may not be applied to nominal data, because there is no implied order in nominal data. Green is not greater than yellow, and red is not less than blue. Only the set algebra operations equal to (=) and not equal to (< >) apply to these nominal variables. All set algebra operations may be applied to ordinal data, and all are often applied to interval/ratio data.

Boolean Algebra

Boolean algebra uses the conditions OR, AND, and NOT to select features. Boolean expressions are most often used to combine set algebra conditions and create compound selections. The Boolean expression consists of a set of Boolean operators, variables, and perhaps constants or scalar values.

Boolean expressions are evaluated by assigning an outcome, true or false, to each condition. Figure 9-5 shows three examples of Boolean expressions. The first is an expression using a Boolean AND, with two arguments for the expression. The first argument specifies a condition on a variable named area, and the second argument a condition on a variable named farm_income. Features are selected if they satisfy both arguments, that is, if their area is larger than 100,000 AND farm_income is less than 10 billion.

Expression 2 in Figure 9-5 illustrates a Boolean NOT expression. This condition specifies that all features with a variable state which is not equal to Texas will return a true value, and hence be selected. NOT is also often known as the negation operator. This is because we might interpret the application of a NOT operation as exchang-

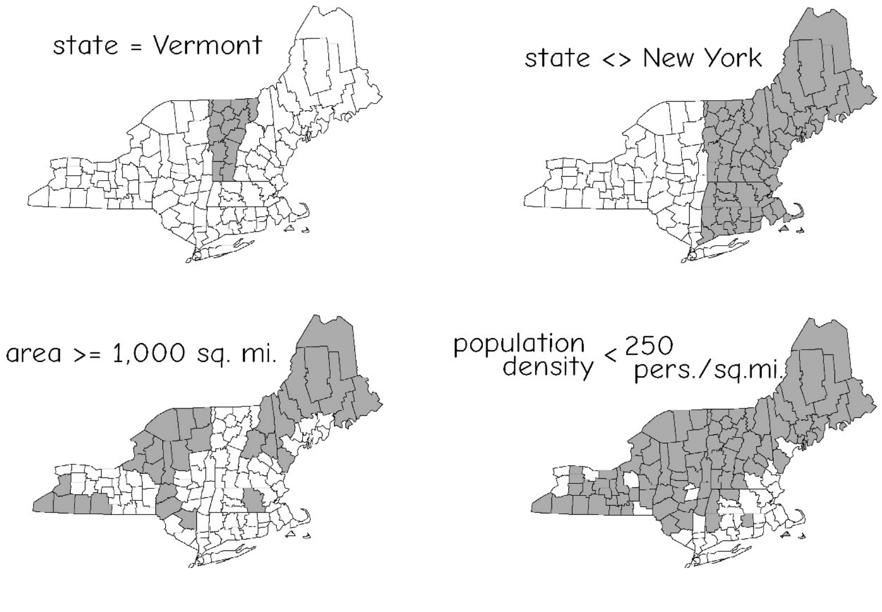


Figure 9-4: Examples of expressions in set algebra and their outcome. Selected features are shaded.

Boolean expressions

```
1. (area > 100,000)
   AND
   (farm_income < 10 billion)

2. NOT ( state = Texas )

3. [ ( rainfall > 1,000 )
     AND
     ( taxes = low) ]
    OR
   [ ( house_cost < 65,000)
     AND
     NOT (crime = high) ]
```

Figure 9-5: Examples of Boolean expressions.

ing the selected set for the unselected set. The argument of expression 2 in Figure 9-5 is itself a set algebra expression. When applied to a set of features, this expression will select all features for which the variable state is equal to the value Texas. The NOT operation reverses this, and selects all features for which the variable state is not equal to Texas.

The third expression in Figure 9-5 shows a compound Boolean expression, combining four set algebra expressions with AND, OR, and NOT. This example shows what might be a naive attempt to select areas for retirement. Our grandparent is interested in selecting areas that have high rainfall and low taxes (a gardener on a fixed income), or low housing cost and low crime.

The spatial outcomes of specific Boolean expressions are shown in Figure 9-6. The figure shows three overlapping circular regions, labeled A, B, and C. Areas may fall in more than one region; for example, the center, where all three regions overlap, is in A, B, and C. As shown in the figure, Boolean AND, OR, or NOT may be used to select any combination or portions of these regions.

OR conditions return a value of true if either argument is true. Areas in either region A or region B are selected at the top center of Figure 9-6. AND requires the conditions on both sides of the operation be met; an AND operation results in a reduced selection set (top right, Figure 9-6). NOT is the negation operator, it flips the effect of the previous operations; it turns true to false and false to true. The NOT shown in the lower left portion of Figure 9-6 returns the area that is only in region C. Note that this is the converse, or opposite set that is returned when using the comparative OR, shown in the top center of Figure 9-6. The NOT operation is often applied in combination with the AND Boolean operator, as shown at the bottom center of Figure 9-6. Again, this selects the converse (or com-

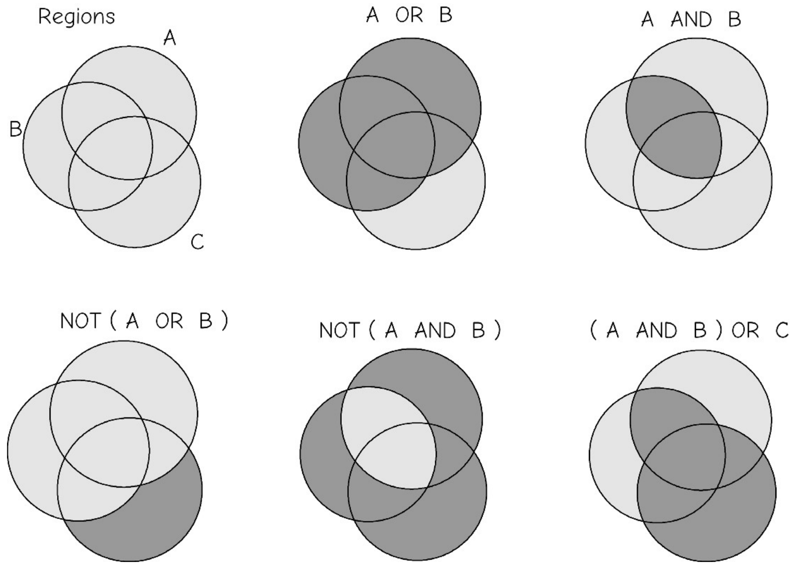


Figure 9-6: Examples of expressions in Boolean algebra, and their outcomes. Subareas of three regions are selected by combining AND, OR, and NOT conditions in Boolean expressions. Any sub-area or group of sub-areas may be selected by the correct Boolean combination.

(County = Rice)
AND
(Wshed = Cannon)

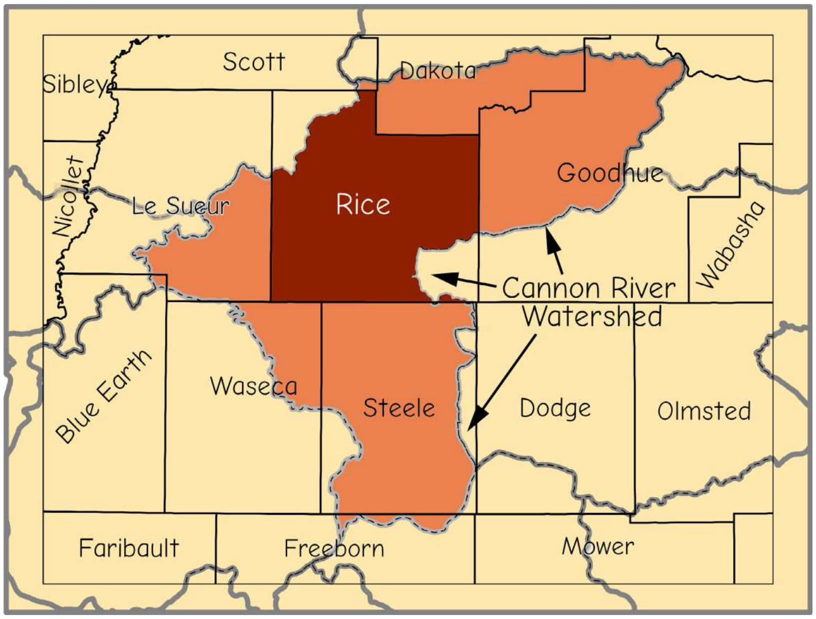


Figure 9-7: An example of a Boolean selection applied to a set of counties and watersheds in the midwestern United States. The mid shaded area is within the Cannon River watershed, a tributary of the Mississippi River, while the darkest shaded area is also within Rice County, Minnesota.

plement) of the corresponding AND. Compare the bottom center selection to the top right selection in Figure 9-6. NOTs, ANDs, and ORs may be further combined to select specific combinations of areas, as shown in the lower right of Figure 9-6.

Note that as with table selection discussed in Chapter 8, the order of application of these Boolean operations is important. In most cases, you will not select the same set when applying the operations in a different order. Parentheses, brackets, or other delimiters should be used to specify the order of application. The expression A AND B OR C will give different results when interpreted as (A AND B) OR C, as shown in Figure 9-6, than when interpreted as A AND (B OR C). Verify this as an exercise. Which areas does the second Boolean expression select?

Figure 9-7 shows a real-world example of a Boolean selection. Counties often must identify areas for treatment, in this case a portion of the Cannon River, a tributary of the Mississippi River, targeted for pollution reduction. Counties are labeled, with boundaries shown as thick solid lines. The Cannon River watershed is shown in darker shades of gray. A Boolean AND operation was applied to a data layer containing both watershed and county boundaries, selecting the areas that are both within the Cannon River watershed and within Rice County.

Spatial Selection Operations

Many spatial operations select sets of features. These operations are applied to a spatial data layer and return a set of features that meet a specified condition. Adjacency and containment are commonly used spatial selection operations.

Adjacency selection operations are used to identify those features that "touch" other features. Features are typically considered to touch when they share a boundary, as when two polygons share an edge. A target or key set of polygon features is identified, and all features that share a boundary with the target features are placed in the selected set.

Figure 9-8a shows an example of a selection based on polygon adjacency. The state of Missouri is shaded on the left side of Figure 9-8a, and states adjacent to Missouri are shaded on the right portion of Figure 9-8a. States are selected because they include a common border with Missouri.

There are many ways the shared border may be detected. With a raster data layer, an exhaustive cell-by-cell comparison may be conducted to identify adjacent pairs with different state values. Vector adjacency may be identified by observing the topological relationships (see Chapter 2 for a discussion of topology). Line and polygon topology typically records the polygon identifiers on each side of a line. All lines with Missouri on one side and a different state on the other side may be flagged, and the list of states adjacent to Missouri extracted.

Adjacency is defined in Figure 9-8a as sharing a boundary for some distance greater than zero. Figure 9-8b shows how a different definition of adjacency may affect selection. The left of Figure 9-8b shows the state of Arizona and a set of adjacent (shaded) western states. By the definition of adjacency used in Figure 9-8a, Arizona and Colorado are not adjacent, because they do not share a boundary along a line segment. Arizona and Colorado share a border at a point, called Four Corners, where they join with Utah and New Mexico. When a different definition of adjacency is used, with a shared node qualifying as adjacent, then Colorado is added to the selected adjacent set (right, Figure 9-8b). This is another illustration of an observation made earlier; there are often several variations of any single spatial

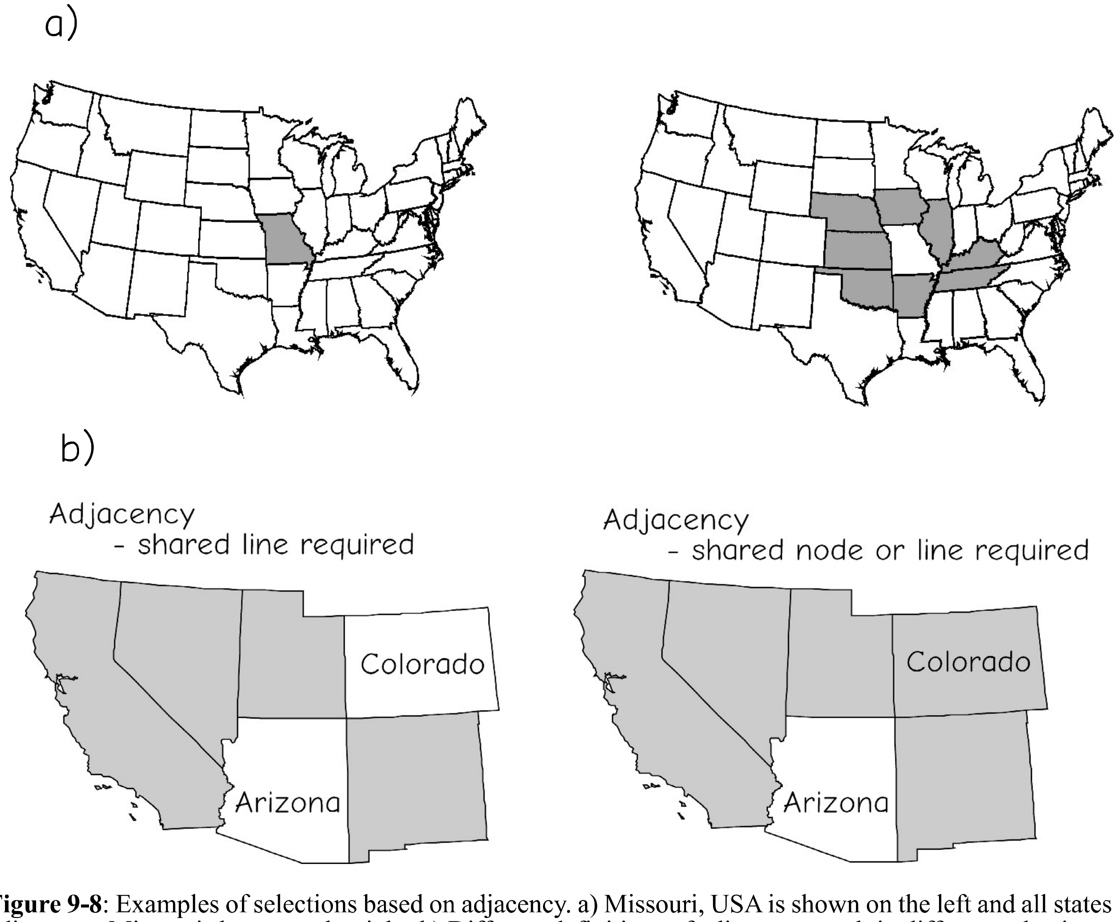


Figure 9-8: Examples of selections based on adjacency. a) Missouri, USA is shown on the left and all states adjacent to Missouri shown on the right. b) Different definitions of adjacency result in different selections. Colorado is not adjacent to Arizona when line adjacency is required (left), but is when node adjacency is accepted (right).

operation. Care must be taken to test the operation under controlled conditions until the specific implementation of a spatial operation is well understood.

Features may be selected based on proximity. Proximity selection typically requires a set of selecting features and a set of target features. All the target features within a specified distance of the selecting features will be chosen, e.g., all weather stations within 60 km of a watershed may be selected to estimate rainfall (Figure 9-9). Selecting only stations within the watershed may provide poor estimates of rainfall near the edge of the polygon, but using all gages is inefficient because information from gages very far away often isn't helpful. An adequate exterior proximity is chosen, usually by visual inspection, previous experience, or preliminary tests, and the proximity selection applied. Proximity is usually calculated implicitly within the operation, in that first the source features and selectable features are provided, a proximity distance and method specified, and on running the operation, the selected set is identified. Less frequently, softwares require a multi-step process in which the source features and selection layers area identified, and a selection layer is created. This created layer usually contains polygons defining the area within the specified distance of the source features. This polygon layer is then used to select features from the target set.

There are several variants of proximity selection. One variant selects all features that are at least partially within a given distance of a set of features. Another variant selects only features that are entirely inside a given distance of the outer boundaries of a set of polygons. A third variant selects only those features that are entirely within a given distance of a set of polygons, but not those that are within the set of polygons themselves. Users should clearly identify the selection tool function and outputs.

Proximity selection, and most selection processes, typically only select features that meet a given set of criteria. The process often does not create a separate, new data layer of the selected features. Selected features are marked on-screen and in the corresponding data table, but still are part of the source data set. There is often an additional export step if these selected features are destined for a new data set. This explicit creation approach is usually

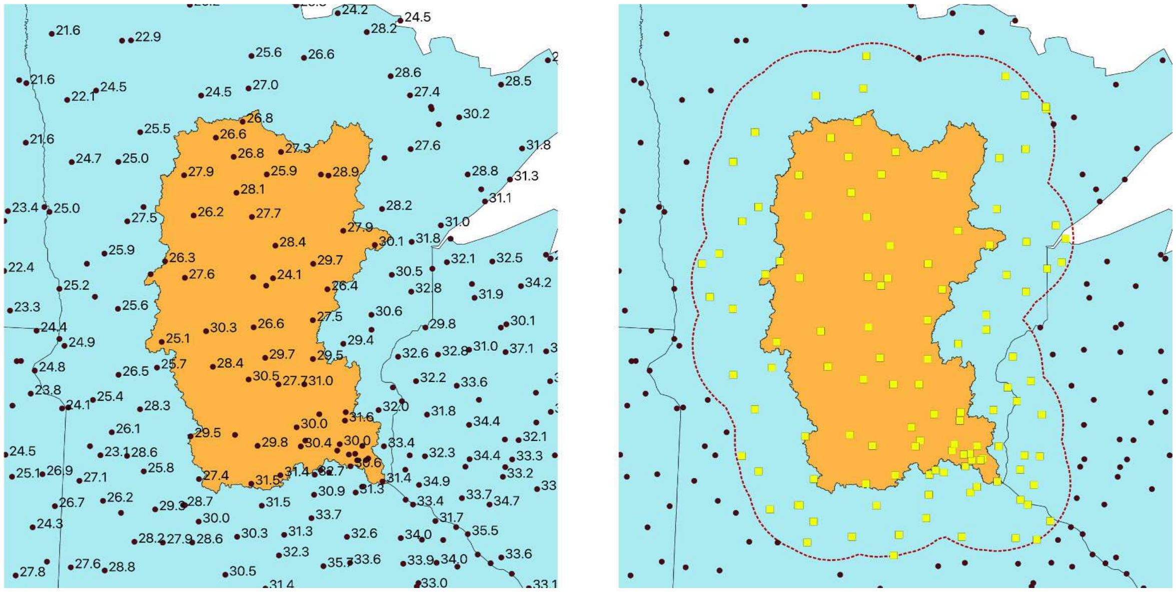


Figure 9-9: An example of a spatial selection based on proximity, where rainfall measured at gages (points, left panel, with mean label) near a watershed are selected. All gages are identified within a 60 km proximity of the watershed boundary (right panel, light squares).

adopted because selection is often a multi-step process, with various different selection tools applied successively to arrive at a target set of features.

We often add indicator or classification variables to our data tables when we have a complex set of selection criteria, particularly when combining both spatial and tabular selections. These indicator variables record the membership of features in groups that match or don't match a set of conditions. Figure 9-10 illustrates a selection for a set of counties that both contain part of a target watershed and hazardous materials (Hazmat) storage sites. A municipality might undertake an upstream inventory if they measure a spike in toxic chemicals in their water supply. Hazmat sites are distributed throughout the state of Georgia. The Altamaha River drains much of the central portion of the state. A spatial selection identifies counties that contain a portion of the watershed, and a table selection, described in Chapter 8, applied to the selected counties reduces the set to those that might be contributing to downstream pollution from Hazmat sites. A column may be added to the county table and values assigned to record this set of potential source counties, to be used in subsequent analysis.

Figure 9-11 shows one flowchart of the geoprocessing analysis in Figure 9-10. While the analysis is relatively simple, and the steps primarily operations on the tabular data, the flowchart shows the data, spatial operations, and order in a succinct manner. It is good practice to flowchart all multi-step spatial analyses, and the value of the flowchart grows with the complexity of the analysis.

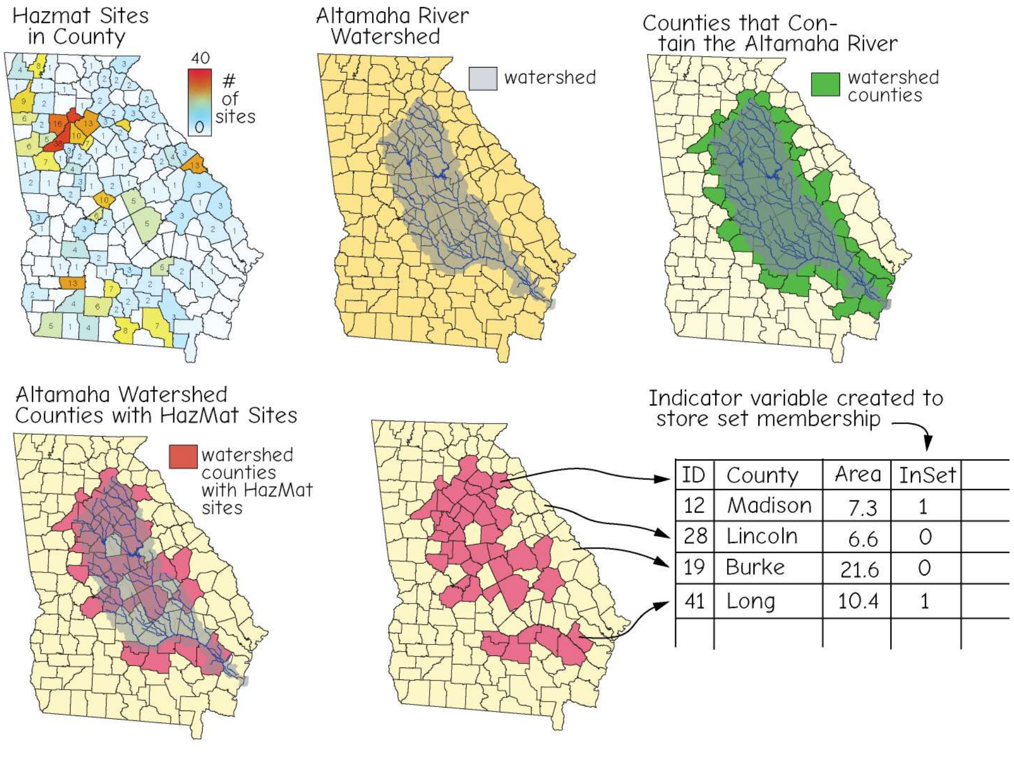


Figure 9-10: An example of indicator variables to record set membership. We wish to identify counties with hazardous materials sites (Hazmat, upper left) that contain part of the Altamaha River watershed (upper center). We may first apply a spatial selection on counties with the watershed boundary (upper right), then a table selection on the coincident counties to identify those with Hazmat sites (lower left). A column, here named InSet, may be added to identify the selected counties in further processing (bottom center and right).

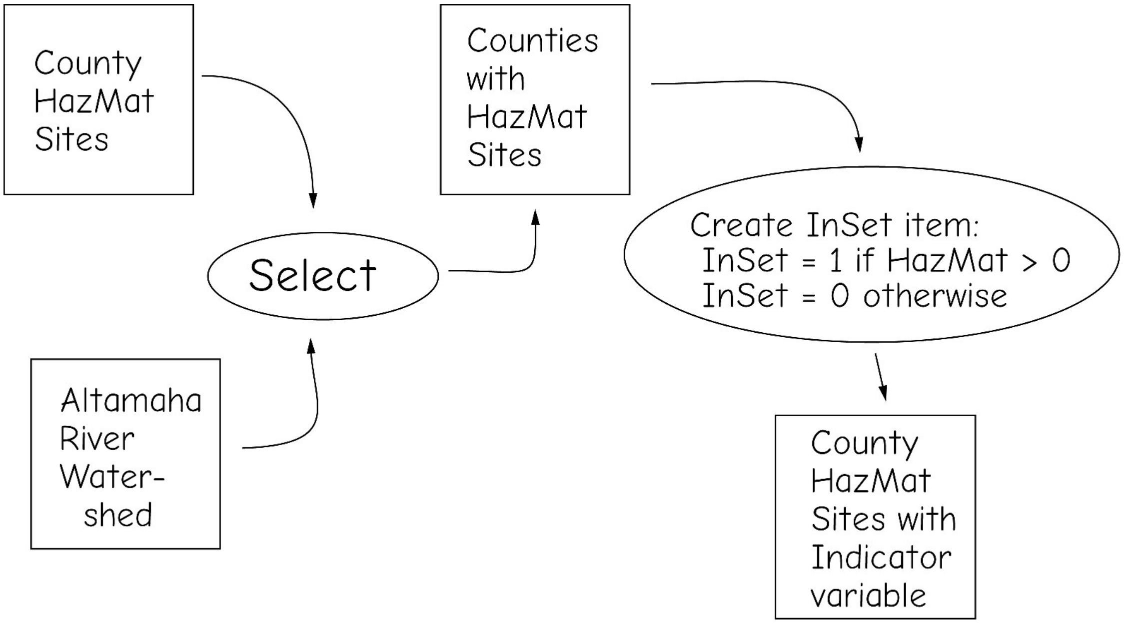


Figure 9-11: An example flowchart for the Altamaha HazMat site analysis. While simple, this shows how spatial and tabular operations might be represented graphically in a diagram.

Caution is helpful when applying subsequent spatial operations to a data set with selected features, e.g., the rainfall gages in the right panel of Figure 9-9. I might recalculate values in the attribute table, copy features, or apply another spatial operation. Some operations by default only act on selected features, while other operations apply to the entire data set. The choices are software-dependent, and so you should consult the documentation or test each new spatial operation when first using it to avoid unintended results

Containment is another spatial selection operation. Containment selection identifies all features that contain or surround a set of target features. For example, the California Department of Transportation may wish to identify all counties, cities, or other governmental bodies that contain some portion of Highway 99, because they wish to improve road safety. A spatial selection may be used to identify these governmental bodies.

Figure 9-12 illustrates a containment selection based on the Mississippi River in North America. We wish to identify states that contain some portion of the river and its tributaries. A query is placed, identifying the features that are contained, here the Mississippi River network, and the target features that may potentially be selected. The target set in this example consists of the lower 48 states of the United States. All states that contain a portion of the Mississippi River or its tributaries are shaded as part of the selected set.

Classification

Classification is a spatial data operation that is often used in conjunction with selection. A classification, also known as a *reclassification* or *recoding*, will categorize geographic objects based on a set of conditions. For example, all the polygons larger than one square mile may be assigned a size value equal to Large, all polygons from 0.1 to 1 square mile may be assigned a size equal to Mid, and all polygons smaller than 0.1 square miles may be assigned a size equal to Small (Figure 9-13). Classifications may add to or modify the attribute data for each geographic object. These modified attributes may in turn be used in further analyses, such as for more complex combinations in additional classification.

Figure 9-12: An example of a selection based on containment. All states containing a portion of the Mississippi River or its tributaries are selected.

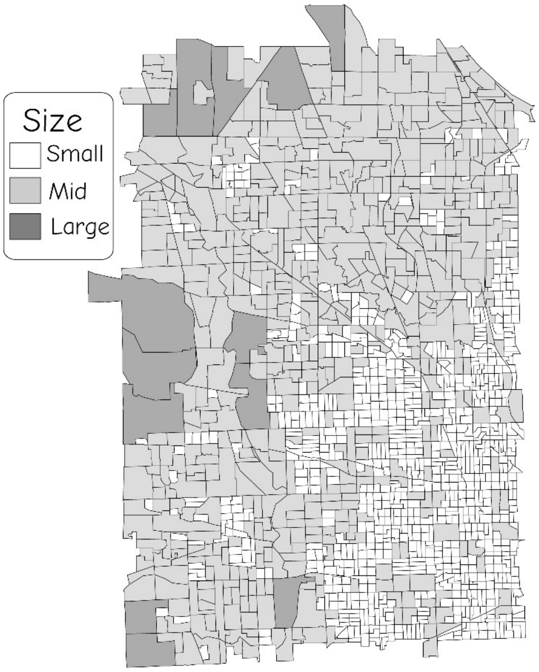


Figure 9-13: Land parcels re-classified by area.

Classification may be used for many other purposes. One common end is to group objects for display or map production. These objects have a common property, and the goal is to display them with a uniform color or symbol so the similar objects are identified as a group. The display color and/or pattern is typically assigned based upon the values of an attribute or attributes. A range of display shades may be chosen, and corresponding values for a specific attribute assigned. The map is then displayed based on this classification.

A classification may be viewed as an assignment of features from an existing set of classes to a new set of classes. We identify features that have a given set of values, for example, parcels that are above a certain size, and assign them all a classification value, in this case the class "large." Parcels in another range of sizes may be assigned different class values, for example, "mid" and "small." The attribute that stores the parcel area is used as a guide to assigning the new class value for size.

The assignment from input attribute values (area) to new class values (here, size) may be defined manually, or the assignment may be defined automatically. For manual classifications, the class transitions are specified entirely by the human analyst.

Classifications are often specified by a table or array. The table identifies the input class or values, as well as the output class for each of this set of input values. Figure 9-14 illustrates the use of a classification table to specify class assignment. Input values of A or B lead to an output class value of 1, an input value of E leads to an output value of 2, and an input value of I leads to an output value of 3. The table provides a complete specification for each classification assignment.

Figure 9-14 illustrates a classification based on a manually defined table. A human analyst specifies the In items for the source data layer via a classification table, as well as the corresponding output value for each In variable. Out values must be specified for each input value or there will be undefined features in the output layer. Manually defining the classification table provides the greatest control over class assignment. Alternatively, classification tables may be automatically assigned, in that a number of classes may be specified and some rule embodied in a computer algorithm used to assign output classes for each of the input classes.

A *binary classification* is perhaps the simplest form of classification. A binary classification places objects into two classes: 0 and 1, true and false, A and B, or some other two-level classification. A set of features is selected and assigned a value, and the complement of the set, all remaining features in the data layer, is assigned the different binary value.

A binary classification is often used to store the results of a complex selection operation. A sequence of Boolean and set algebra expressions may be used to select a set of features. A specific target attribute is identified for the selected set of features.

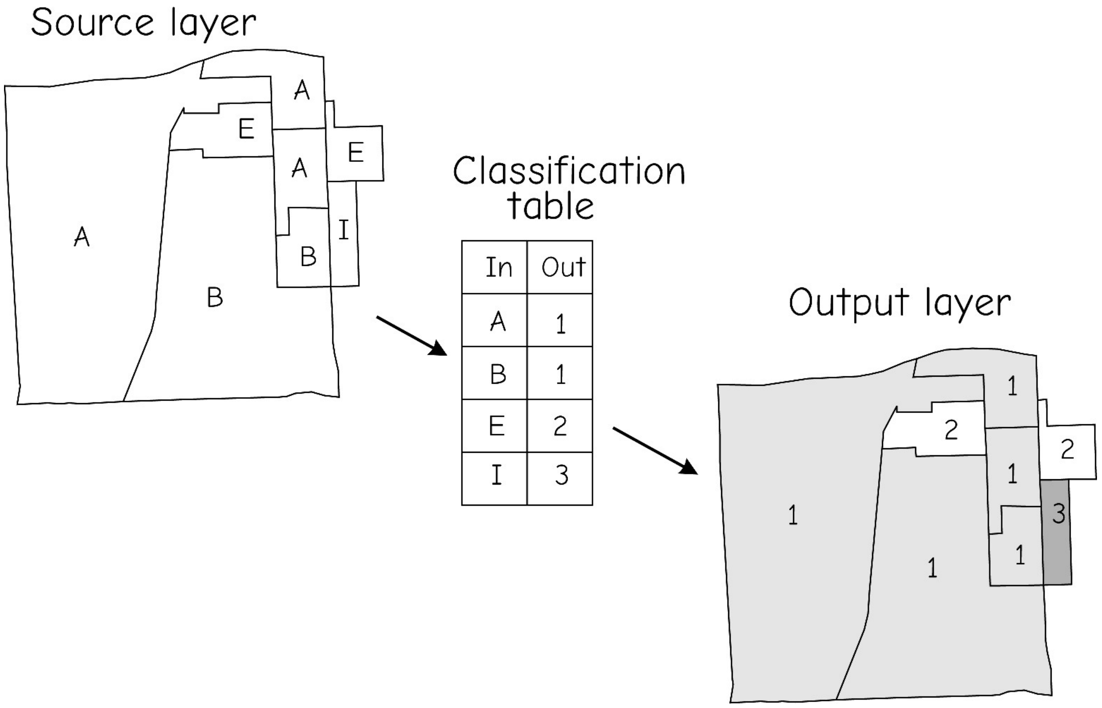


Figure 9-14: The classification of a thematic layer. Values are given to specific attributes in a classification table, which is used, in turn, to assign classes in an output layer.

This target attribute is assigned a unique value. The target attribute is assigned a different value for all unselected features. This creates a column that identifies the selected set; for example, all counties that are small, but with a large population.

For example, we may wish to select states at least partially west of the Mississippi River as an intermediate step in an analysis (Figure 9-15). We may be using this classification in many subsequent spatial operations. Thus, we wish to store this characteristic, whether the state is west or east of the Mississippi River. States are selected based on location and reclassified. We record this classification by creating a new attribute and assigning a binary value to this attribute, 1 for those parcels that satisfy the criteria, and 0 for those that do not (Figure 9-15). The variable is_west records the state location relative to the Mississippi River. Additional selection operations may be applied, and the created binary variable preserves the information generated in the initial selection.

Previous examples have shown vector data, but we may also reclassify raster data. If the input raster values are nominal or ordinal data, the reclassification will look very similar to the vector examples shown in Figure 9-14. A list of input and corresponding outputs are provided, and the reclassification operates on a cell-by-cell basis. When interval/ratio raster data are used as input, then input ranges are required rather than specific input values. This distinction is described in detail in Chapter 10.

Classification table

state name	is_west
Alabama	0
Arizona	1
Arkansas	1
Colorado	1
Connecticut	0
....	
Wyoming	1

Figure 9-15: An example of a binary classification. Features are placed into two classes in a binary classification, west (1) and east (0) of the Mississippi River. The classification table codifies the assignment.

Data-defined Classification

Manually defining the classification table may not always be necessary, and may be tedious or complex. Suppose we wish to assign a set of display colors to a set of elevation values. There may be thousands of distinct elevation values in the data layer, and it would be inconvenient at best to assign each color manually. Data-defined classification methods, where class intervals are automatically derived from rules applied to the input data, are often used in these instances.

An automatic classification uses some rule to specify the input class to output class assignments. The input and output class boundaries are often based on a set of parameters used to guide class definition.

A potential drawback from an automated class assignment stems from our inability to precisely specify class boundaries. A mathematical formula or algorithm defines the class boundaries, and so specific classes of interest may be split. Thus, the analyst sacrifices precise control over class specification when an automated classification is used.

Figure 9-16 describes a data layer we will use to illustrate automatic class assignment. The figure shows a set of "neighborhoods" with populations that range from 0 to 5133. We wish to display the neighborhoods and populations in three distinct classes, high, medium, and low population. High will be shown in black, medium in gray, and low in white. We must decide how to assign the categories — what population levels define high, medium, and low? In many applications, the classification levels are previously defined. There may be an agreed-upon standard for high population, and we would simply use this level. However, in many instances the classes are not defined, and we must choose them.

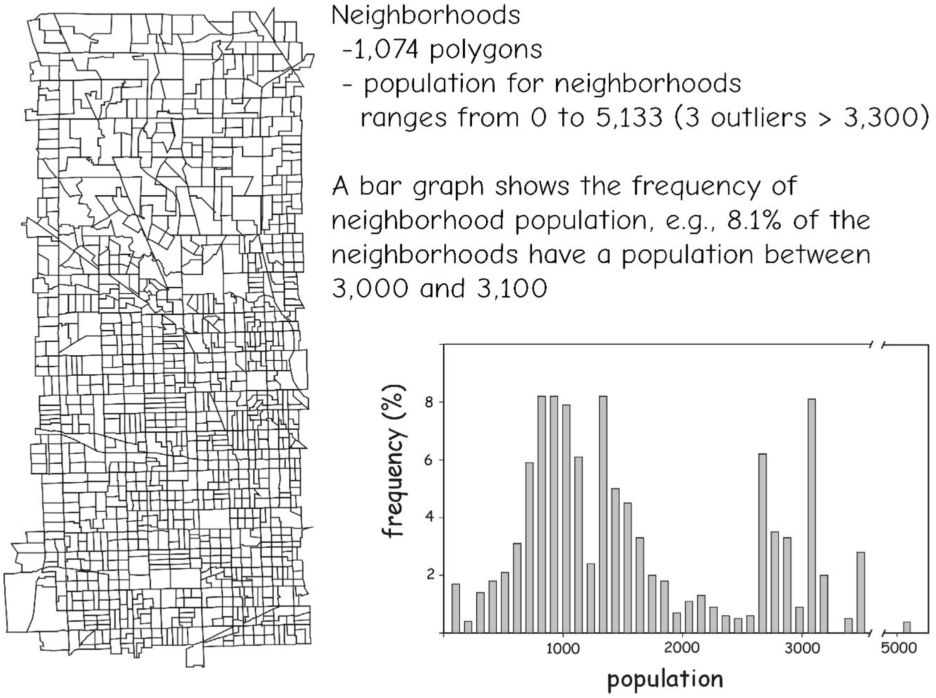


Figure 9-16: Neighborhood polygons and population levels used in subsequent examples of classification assignment. The populations for these 1,074 neighborhoods range from 0 to 5,133. The histogram at the lower right shows the frequency distribution. Note that there is a break in the chart between 3,500 and

Figure 9-16 includes a bar graph depicting the population frequency distribution; this type of bar graph is commonly called a histogram. The frequency histogram shows the number of neighborhoods that are found in each bar (or "bin") of a set of very narrow population categories. For example, we may count the number of neighborhoods that have a population between 3,000 to 3,100. Approximately 8.1% of the neighborhoods have a population in this range, so a vertical bar corresponding to 8.1 units high is plotted. We count and plot the histogram values for each of our narrow categories (e.g., the number from 0 to 100, from 100 to 200, from 200 to 300), until the highest population value is plotted.

Our primary decision in class assignment is where to place the class boundaries. Should we place the boundary between the low and medium population classes at 1,000, or at 1,200? Where should the boundary between medium and high population classes be placed? The location of the class boundaries will change the appearance of the map, and also the resulting classification.

One common method for automatic classification specifies the number of output classes and requests equal-interval classes over the range of input values. This *equal-interval* classification simply subtracts the lowest value of the classification variable from the highest value, and defines equal-width boundaries to fit the desired number of classes into the range.

Figure 9-17 illustrates an equal-interval classification for the population variable. Three classes assigned over the range of 0 to 5,133 are specified. Each interval is approximately one-third of this range. This range is evenly divided by 1,711. The small

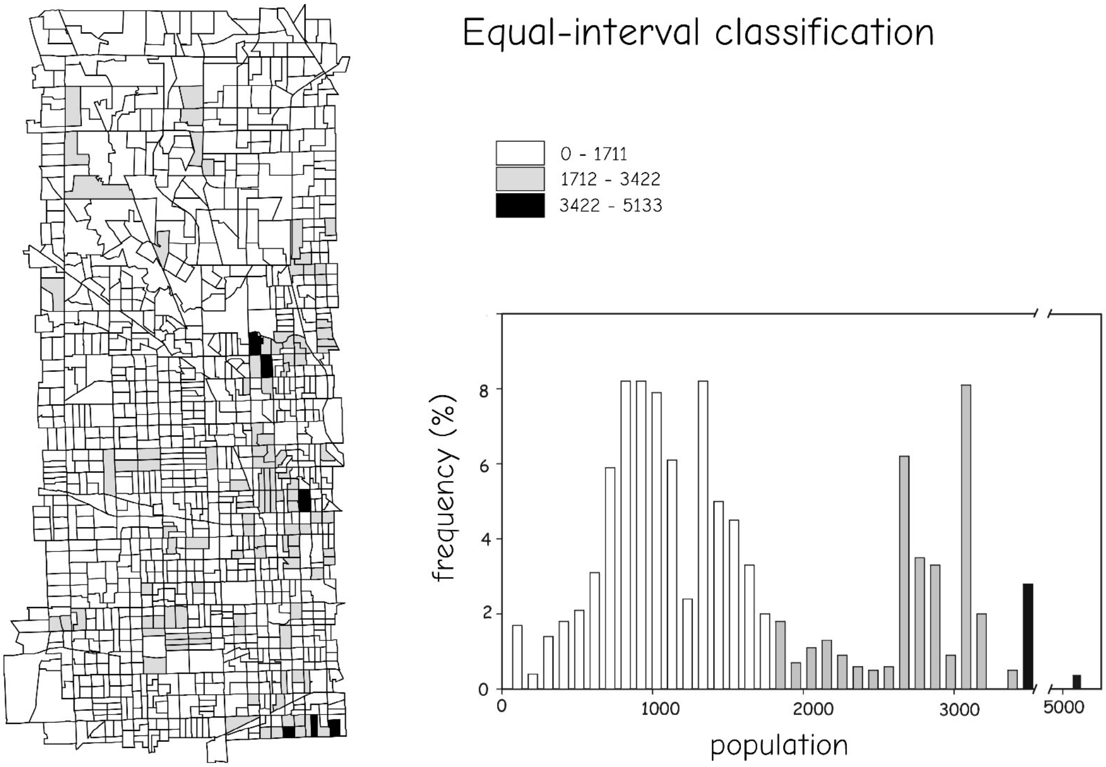


Figure 9-17: An equal-interval classification. The range 0 to 5,133 is split into three equal parts. Colors are assigned as shown in the map of the layer (left), and in the frequency plot (right). Note the relatively few polygons assigned to the high classes in black. A few neighborhoods with populations near 5,000 shift the class boundaries upward.

class extends from 0 to 1,711, the medium class from 1,712 to 3,422, and the large class from 3,423 to 5,133. Population categories are shown colored accordingly on the map and the bar graph, with the small (white), medium (gray), and large (black) classes shown.

Note that the low population class shown in white dominates the map; most of the neighborhoods fall in this population class. This often happens when there are features that have values much higher than the norm. There are a few neighborhoods with populations above 5,000 (to the right of the break in the population axis of the bar graph), while most neighborhoods have populations below 3,000. The outliers shift the class boundaries to higher values, 1711 and 3,422, resulting in most neighborhoods falling in the small population category.

Another common method for class assignment results in an *equal-area* classification (Figure 9-18). Class boundaries are defined to place an equal proportion of the study area into each of a specified number of classes. This usually leads to a visually balanced map because all classes have approximately equal extents. Equal-area classes are often desirable, for example, when resources need to be distributed over equal areas, or when equally sized overlapping sales territories may be specified.

Class width may change considerably with an equal-area classification. An equal-area classification sets class boundaries so that each class covers approximately the same area. A class may consist of a few or even one large polygon. This results in a small range for the large polygon classes. Classes also tend to have a narrow range of values near the peaks in the histogram. Many polygons are represented at the his-

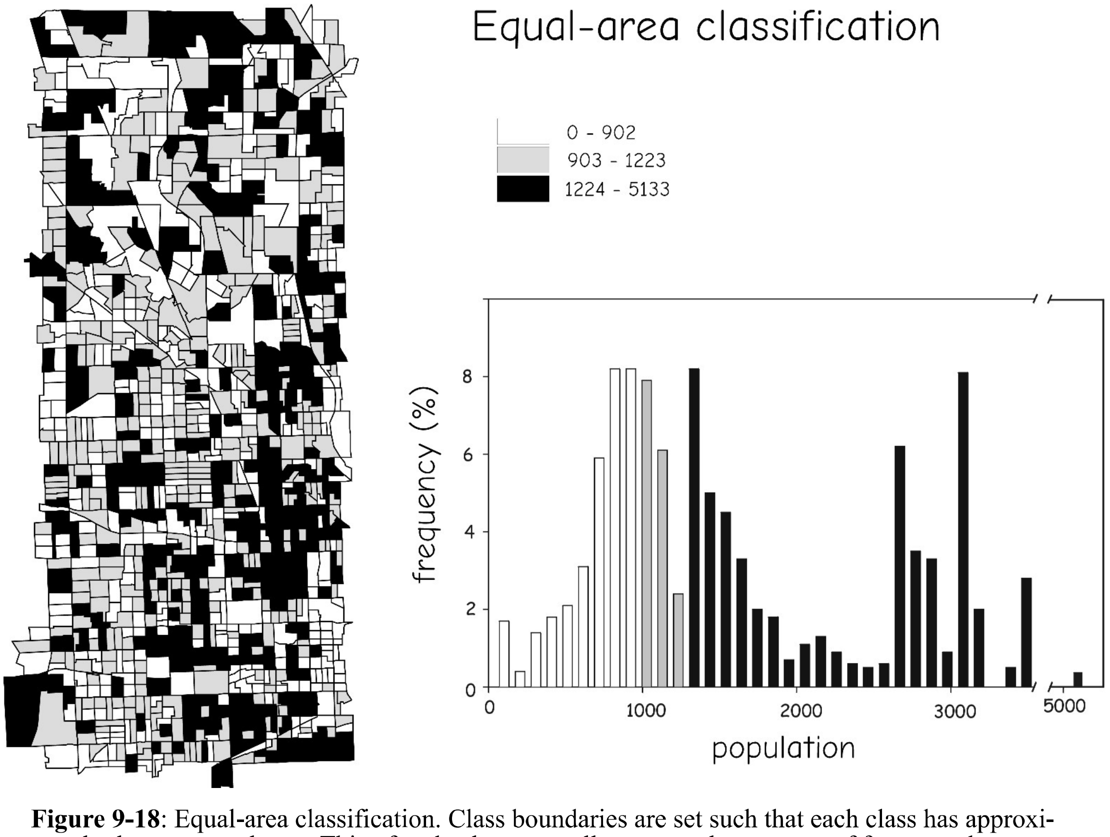


Figure 9-18: Equal-area classification. Class boundaries are set such that each class has approximately the same total area. This often leads to a smaller range when groups of frequent classes are found. In this instance the medium class spans a small range, from 903 to 1,223, while the high population class spans a range that is almost 10 times broader, from 1,224 to 5,133.

togram peaks, and so these may correspond to large areas. Both of these effects are illustrated in Figure 9-18. The middle class of the equal-area classification occurs at population values between 903 and 1223. This range of populations is near the peak in the frequency histogram, and these population levels are associated with larger polygons. This middle class spans a range of approximately 300 population units, while the small and large classes span near 900 and 4,000 population units, respectively.

Equal-area assignments may be highly skewed when there are a few polygons with large areas, and these polygons have similar values. Although not occurring in our example, there may be a relationship between the population and area for a few neighborhoods. Suppose in a data set similar to ours there is one very large neighborhood dominated by large parks. This neighborhood has both the lowest populations and largest area. An equal-area classification may place this neighborhood in its own class. If a large parcel also occurs with high population levels, we may get three classes: one parcel in the small class, one parcel in the high population class, and all the remaining parcels in the medium population class. While most equal-area classifications are not this extreme, unique parcels may strongly affect class ranges in an equal-area classification.

We will cover a final method for automated classification, a method based on *natural breaks*, or gaps, in the data (Figure 9-19). Natural breaks classification looks for "obvious" breaks. It attempts to identify naturally occurring clusters of data; not clusters based on the spatial relationships, but rather clusters based on an ordering variable.

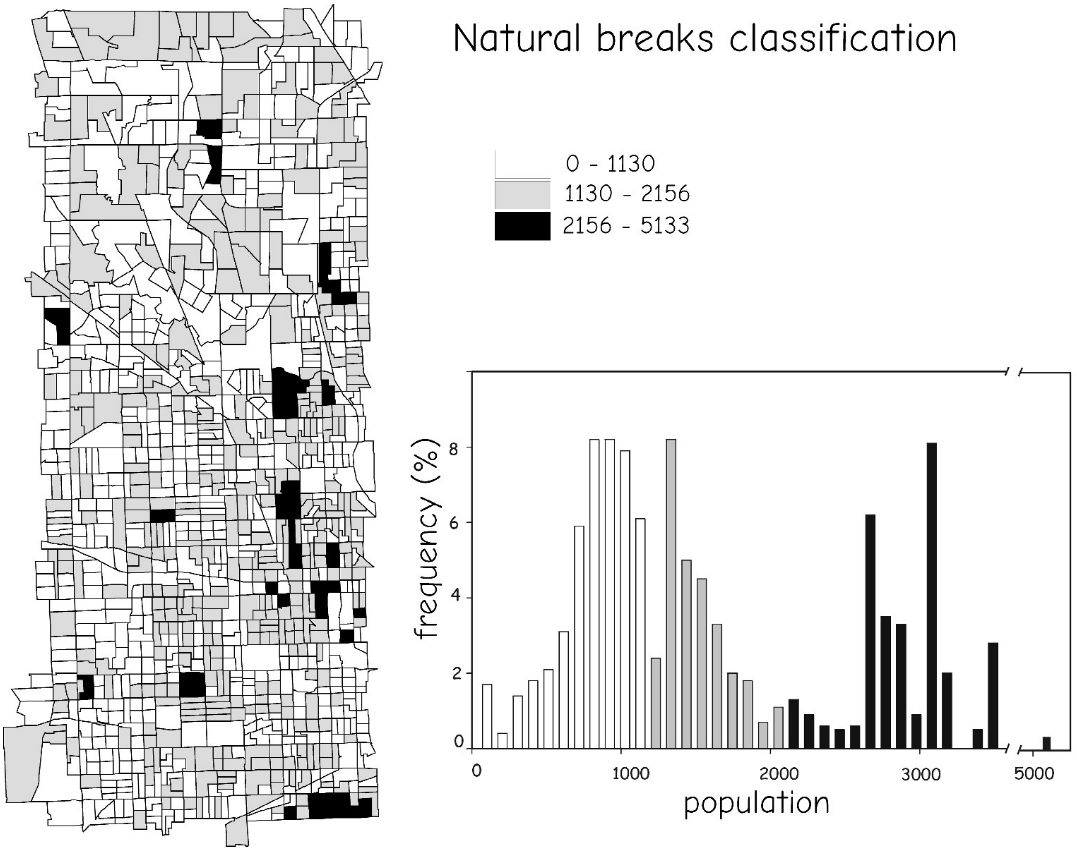


Figure 9-19: Natural breaks classification. Boundaries between classes are placed where natural gaps or low points occur in the frequency distribution.

There are various methods used to identify natural breaks. Large gaps in an ordered list of values are one common method. The values are listed from lowest to highest, and the largest gaps in values selected. Barring gaps, low points in the frequency histogram may be identified. There is usually an effort to balance the need for relatively wide and evenly distributed classes and the search for natural gaps. Many narrow classes and one large class may not be acceptable in many instances, and there may be cases where the specified number of gaps does not occur in the data histogram. More classes may be requested than obvious gaps, so some natural break methods include an alternative method, for example, equally spaced intervals, for portions of the histogram where no natural gaps occur.

Figure 9-19 illustrates a natural break classification. Two breaks are evident in the histogram, one near 1300 and one near 2200 persons per neighborhood. Small, medium, and large populations are assigned at these junctures.

Figure 9-17 through Figure 9-19 illustrate an important point: you must be careful when interpreting class maps, because the apparent relative importance of categories may be manipulated by altering the starting and ending values in each class. Figure 9-17 suggests most neighborhoods are low population, Figure 9-18 suggests that high population neighborhoods cover the largest areas and that they are well mixed with areas of lower population, while Figure 9-19 indicates the area is dominated by low and medium population neighborhoods. Precisely because there are no objectively defined population boundaries, we have great flexibility in manipulating the impression we create. The legend in class maps should be scrutinized, and the range between class boundaries noted. .

The Modifiable Areal Unit Problem

When there are no objectively recognizable categories, polygons may be reclassified and grouped in many ways.The aggregate values for the classified polygons, such as class population, age, and income, will depend on the size, shape, and location of the aggregated polygons (Figure 9-20). This general phenomenon, known as the modifiable areal unit problem or MAUP, has been exploited by politicians to redraw political boundaries to one party or another's — but generally not the country's — advantage. The process of aggregating neighborhoods to create majority blocks for political advantage was named gerrymandering after Massachusetts governor Elbridge Gerry, when he crafted a political district shaped like a salamander.

There are two primary characteristics of the MAUP that may be manipulated to affect aggregate polygon values. The first is the *zoning effect*, that aggregate statistics may change by the shape of the units, and the second is the *size effect*, that aggregate statistics may change with the area of the units. For example, the mean income of a unit will change when the boundaries of a unit change, either because of a change in zone or a change in size.

MAUP effects may substantially influence the values for each unit, and hence subsequent analysis. Openshaw and Taylor published results in 1979 that illustrate MAUP dependencies particularly well. They analyzed the percentage of elderly voters and the number of Republican voters for the 99 counties in Iowa. They showed that the correlation between the elderly voters and Republican voter numbers ranged from 0.98 to -0.81 by varying the scale and aggregation units that grouped counties. Additional work has shown that multivariate statistical models based on aggregate data are similarly dependent on the aggregation units, leading to contradictory results predictions.

Numerous studies of the MAUP have shown how to identify and/or reduce the pri-

mary negative impacts of the zoning and size effects. The primary recommendation is to work with the basic units of measure. In our census example, this would be to collect and maintain information on the individual person or household. This is often not possible; for example, aggregation is specifically required to maintain the anonymity of the census respondents. However, many efforts allow recording and maintaining data on the primary units within a GIS framework, and this should be implemented when possible.

A second way to address the MAUP is based on optimal zoning. Zones are designed to maximize variation between zones while minimizing variation within zones. Optimal zones are difficult to define for more than one variable, because variables often do not change in concert. For example, an optimal set of zones for determining traffic densities may not be an optimal for average age. Old people are no more nor less likely to live near busy or low traffic roads. Optimum zoning approaches are best applied when interest in one variable predominates.

Another approach to solving the MAUP involves conducting a set of sensitivity analyses. Units are aggregated and rezoned across a range of sizes and shapes and the analyses performed for each set. Changes in the results are observed, and the sensitivity to zone boundaries and sizes noted. These tests may identify the relative sensitivities of different variables to size and zoning effects. Robust results may be identified, for example, average age may not change over a range of sizes and unit combinations, yet may change substantially over a narrow range of sizes in some areas. This approach requires many computations, because it uses replicated runs for each set of variables, zone levels, and shapes. This often overwhelms the available computing resources for many problems and agencies.

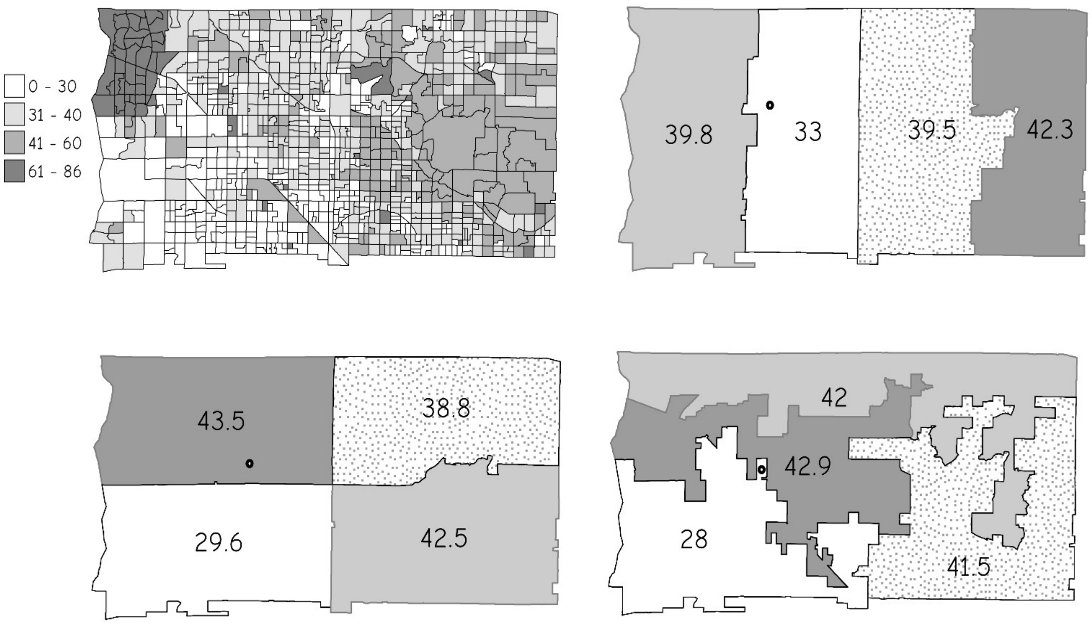


Figure 9-20: An example of the modifiable areal unit problem (MAUP). Census blocks (upper left) have been aggregated in various ways to produce units that show different mean population age. All units are approximately equal in size. Note that the number of zones with a mean age over 40 can be one (upper right), two (lower left), or three (lower right), and that an individual block (small circle, upper left quadrant) may be in a polygon with a mean age in the 20s, 30s, or 40s, depending on unit shape.

Dissolve

A *dissolve* function is primarily used to combine similar features within a data layer. Adjacent polygons may have identical values for an attribute. For example, a wetlands data layer may specify polygons with several subclasses, such as wooded wetlands, herbaceous wetlands, or open water. If an analysis requires we identify only the wetland areas vs. the upland areas, then we may wish to dissolve all boundaries between adjacent wetlands. We are only interested in preserving the wetland/upland boundaries.

Dissolve operations are usually applied based on a specific "dissolve" attribute associated with each feature. A value or set of values is identified that belongs in the same grouping. Each line that serves as a boundary between two polygon features is assessed. The values for the dissolve attribute are compared across the boundary line. If the values are the same, the boundary line is removed, or dissolved away. If the values for the dissolve attribute differ across the boundary, the boundary line is left intact.

Figure 9-21 illustrates the dissolve operation that produces a binary classification. This classification places each state of the contiguous United States into one of two categories, those entirely west of the Mississippi River (1) and those east of the Mississippi River (0). The attribute named is_west contains values indicating location. A dissolve operation applied on the variable is_west removes all state boundaries between similar states. This reduces the set from 48 polygons to two polygons.

Dissolve table

state name	is_west	dissolve value
Alabama	0	E
Arizona	1	W
Arkansas	1	W
Colorado	1	W
Connecticut	0	E
....	....	....
Wyoming	1	W

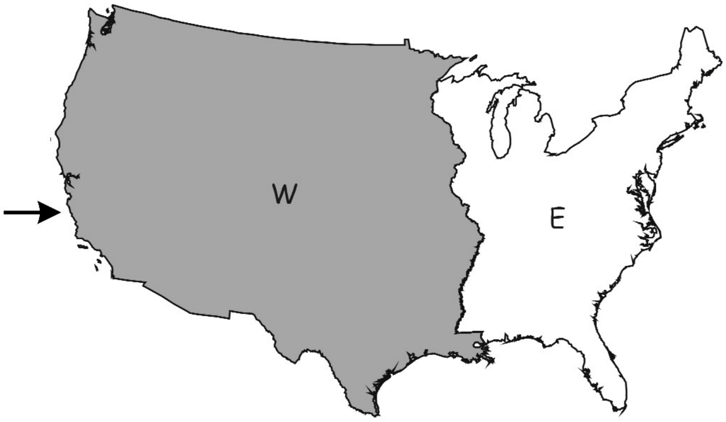


Figure 9-21: An illustration of a dissolve operation. Boundaries are removed when they separate states with the same value for the dissolve attribute is_west.

Dissolve operations are often needed prior to applying an area-based selection in spatial analysis. For example, we may wish to select areas from the natural breaks classification shown on the left of Figure 9-22. We seek polygons that are greater than 3 mi^2 in area and have a medium population. The polygons may be composed of multiple neighborhoods. We typically must dissolve the boundaries between adjacent, medium-sized neighborhoods prior to applying the size test. Otherwise, two adjacent, medium population neighborhoods may be discarded because both cover approximately 2 mi^2. Their total area is 4 mi^2, above the specified threshold, yet they will not be selected unless a dissolve is applied first.

Dissolves are also helpful in removing unneeded information. After the classification into small, medium, and large size classes, many boundaries may become redundant. Unneeded boundaries may inflate storage and slow processing. A dissolve has the advantage of removing unneeded geographic and tabular data, thereby simplifying data, improving processing speed, and reducing data volumes.

Figure 9-22 illustrates the space saving and complexity reduction common when applying a dissolve function. The number of polygons is reduced approximately ninefold by the dissolve, from 1,074 on the left to 119 polygons on the right of Figure 9-22.

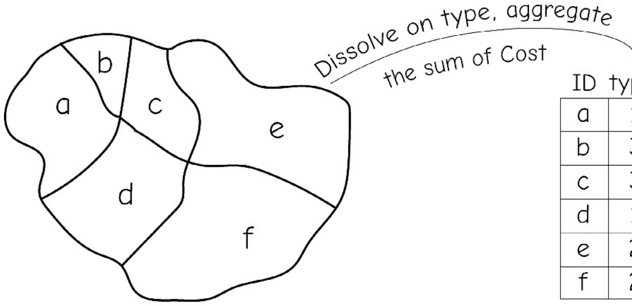
Figure 9-22: An example of a dissolve operation. Note the removal of lines separating polygons of the same size class. This greatly reduces the number of polygons.

Attribute Aggregation in a Dissolve Operation

We often wish to transfer information in the attribute table when applying a dissolve function. For example, we may wish to find the total population of a set of neighborhoods that are within walking distance of a set of bus stops. Each neighborhood polygon contains a population count attribute, and we wish to sum the values across the neighborhoods that correspond to each bus stop. The new dissolved polygons, representing the neighborhoods closest to each bus stop, will contain a summed population variable. We might then do further analysis to identify areas where new bus stops might be needed, or to recommend a change in bus frequency.

Aggregation functions allow us to preserve information in a dissolve operation. Typically we may sum, average, calculate the range, maximum, minimum, or other common statistics, and assign these to the output polygons. The functions first identify the adjacent polygons that will be combined to form the new polygons, and then apply the specified operation to target attribute variables. Figure 9-23 diagrams a dissolve that sums cost across input polygons. Adjacent polygons of the same type are combined, and the values of the component polygons summed. Different analyses might require different aggregation statistics, e.g., average, maximum, or range (Figure 9-24).

Some of the variables might be nonsensical as inputs or as outputs, so care must be taken when aggregating during a dissolve. This is particularly true for area averaged values, or for categorical or ordi-



ID	type	Cost	Virus rating	Lease year	Stand age	Exmp. status
a	1	1,000	0	1979	49	0
b	3	6,000	5	2011	106	0
c	3	3,000	2	1957	24	1
d	1	2,000	9	1989	83	0
e	2	7,000	0	1959	19	1
f	2	9,000	1	1991	206	1

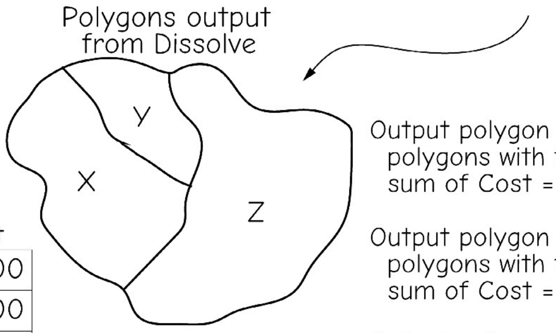


Table output from Dissolve

ID	type	Cost
X	1	3,000
Y	3	9,000
Z	2	16,000

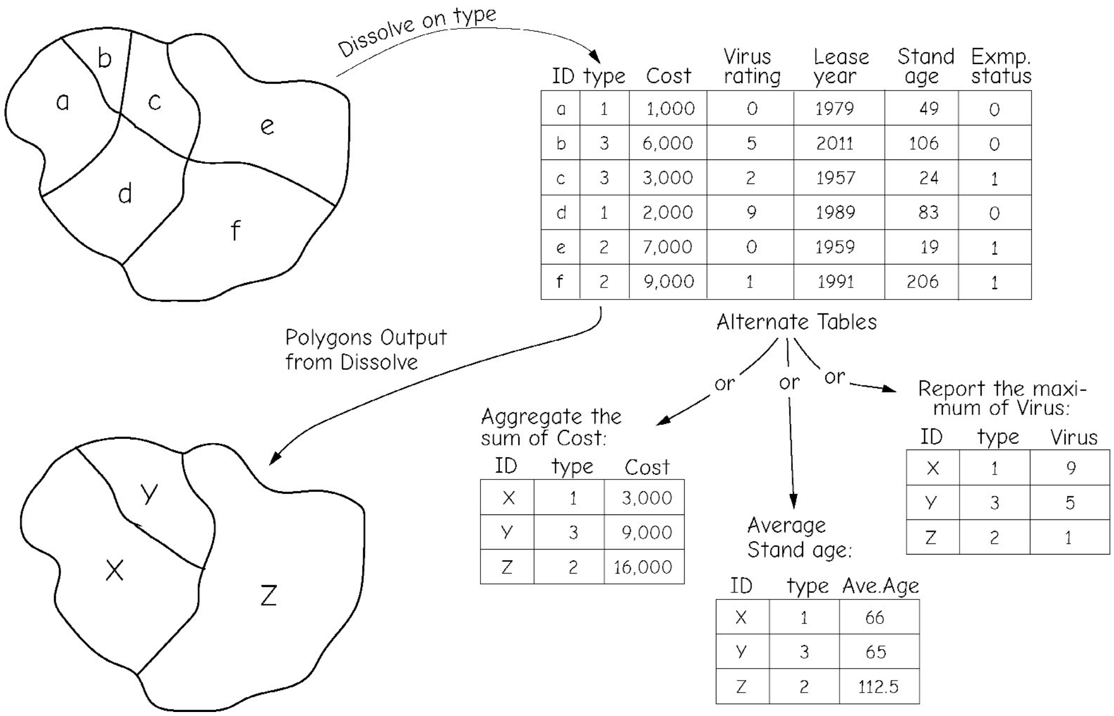
Output polygon X takes values from input polygons with type=1, rows a and d above, sum of Cost = 1,000 + 2,000 = 3,000

Output polygon Y takes values from input polygons with type=2, rows b and c above, sum of Cost = 6,000 + 3,000 = 9,000

Output polygon Z takes values from input polygons with type=3, rows e and f above, sum of Cost = 7,000 + 9,000 = 16,000

Figure 9-23: An example aggregation during a dissolve operation. Adjacent polygons of the same type are combined. For each combination, the input polygons costs are summed and this sum assigned to a cost variable for the aggregated polygon in the output data layer.

nal values as inputs. If I average the average population of two input polygons, I will get an erroneous average for the output polygon, except in the rare case when both polygons have the same area. Proof of this is left to the reader. Aggregates of categorical or ordinal variables also require caution, as the average or sum of ordinal or category values has meaning in relatively few applications.



ID	type	Cost	Virus rating	Lease year	Stand age	Exmp. status
a	1	1,000	0	1979	49	0
b	3	6,000	5	2011	106	0
c	3	3,000	2	1957	24	1
d	1	2,000	9	1989	83	0
e	2	7,000	0	1959	19	1
f	2	9,000	1	1991	206	1

ID	type	Cost
X	1	3,000
Y	3	9,000
Z	2	16,000

ID	type	Ave.Age
X	1	66
Y	3	65
Z	2	112.5

ID	type	Virus
X	1	9
Y	3	5
Z	2	1

Figure 9-24: Different aggregation operations may be applied, and appropriate, for different input variables. Here are examples of sum, average, and maximum operations applied during a dissolve operation.

Proximity Functions and Buffering

Proximity functions or *operations* are among the most powerful and common spatial analysis tools. Many important questions hinge on proximity, the distance between features of interest. How close are schools to an oil refinery, what neighborhoods are far from convenience stores, and which homes will be affected by an increase in freeway noise? Many questions regarding proximity are answered through spatial analyses in a GIS. Here we focus on proximity functions that create new features and layers, rather than proximity selection, described earlier.

Proximity functions modify existing features or create new features that depend in some way on distance. For example, one simple proximity function creates a raster of the minimum distance from a set of features (Figure 9-25). The figure shows a distance function applied to water holes in a wildlife reserve. Water is a crucial resource for nearly all animals, and the reserve managers may wish to ensure that most of the area is within a short distance of water. In this instance point features are entered that represent the location of permanent water. Water holes are represented by individual points, and rivers by a group of points set along the river course. A proximity function calculates the distance to all water points for each raster cell. The minimum distance is selected and placed in an output raster data layer (Figure 9-25). The distance function creates a mosaic of what appear to be overlapping circles. Although the shading scheme shows apparently abrupt transitions, the raster cells contain a smooth gradient in distance away from each water feature.

Distance values are most often calculated based on the Pythagorean formula (Figure 9-26). These values are typically calculated from cell center to cell center when applied to a raster data set. Although any distance is possible, the distances between adjacent cells change in discrete intervals related to the cell size. Note that distances are not restricted to even multiples of the cell size, because distances measured on diagonal angles are not even multiples of the cell dimension. There may

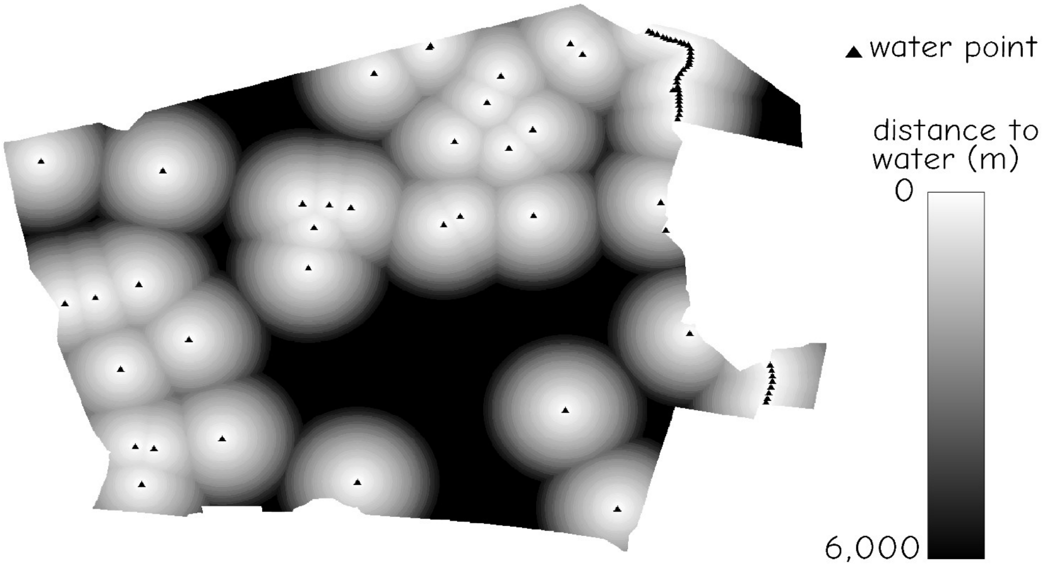


Figure 9-25: An example of a distance function. This distance function is applied to a point data layer and creates a raster data layer. The raster layer contains the distance to the nearest water feature.

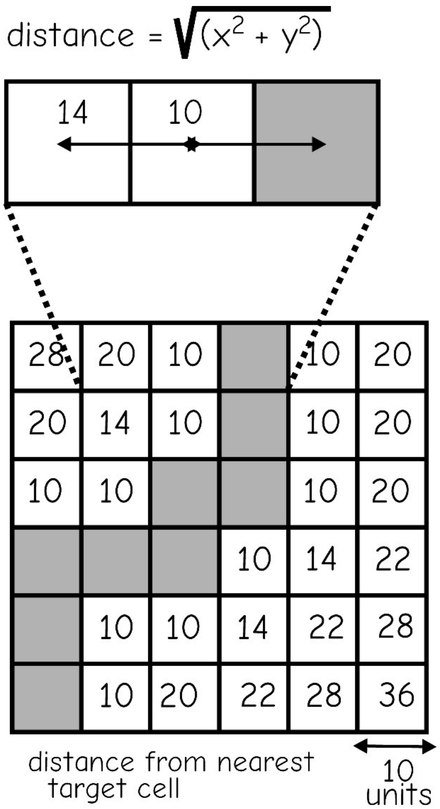


Figure 9-26: A distance function applied to a raster data set.

be no cells that are exactly some fixed distance away from the target features; however, there may be many cells less than or greater than that fixed distance.

Buffers

Buffering is one of the most commonly used proximity functions. A *buffer* is a region that is less than or equal to a specified distance from one or more features (Figure 9-27). Buffers may be determined for point, line, or area features, and for raster or vector data. Buffering is the process of creating buffers. Buffers typically identify areas that are "outside" some given threshold distance compared to those "inside" some threshold distance.

Buffers are used often because many spatial analyses are concerned with distance constraints. For example, emergency planners might wish to know which schools are within 1.5 kilometers of an earthquake fault, a park planner may wish to identify all lands more than 10 kilometers from the nearest highway, or a business owner may wish to identify all potential customers within a given radius of her store. All these questions may be answered with the appropriate use of buffering.

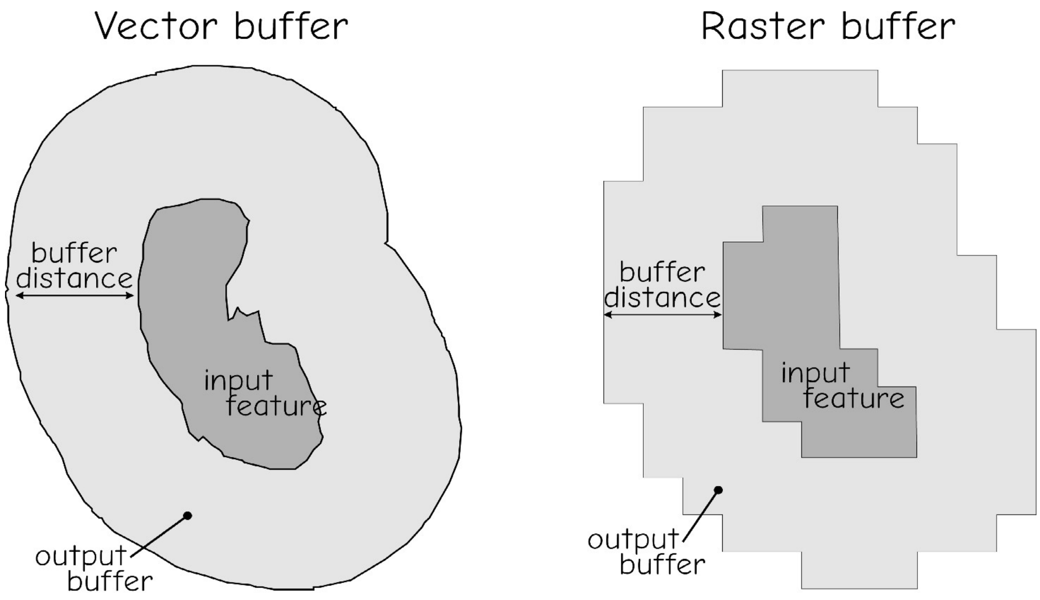


Figure 9-27: Examples of vector and raster buffers derived from polygonal features. A buffer is defined by those areas that are within some buffer distance from the input features.

Raster Buffers

Buffer operations on raster data entail calculating the distance from each source cell center to all other cell centers. Output cells are assigned an in value whenever the cell-to-cell distance is less than the specified buffer distance. Those cells that are further than the buffer distance are assigned an out value (Figure 9-28).

Raster buffers combine a minimum distance function and a binary classification function. A minimum distance function calculates the shortest distance from a set of target features and stores this distance in a raster data layer. The binary classification function splits the raster cells into two classes: those with a distance greater than the threshold value, and those with a distance less than or equal to a threshold value.

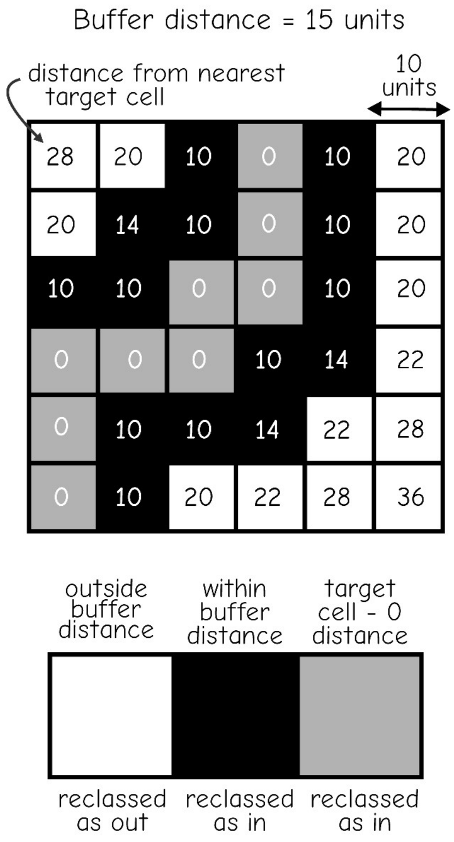


Figure 9-28: Raster buffering as a combination of distance and classification. Here, cells less than 15 units from the target cells are identified.

Buffering with raster data may produce a "stair-step" boundary, because the distance from features is measured between cell centers. When the buffer distance runs parallel and near a set of cell boundaries, the buffer boundary may "jump" from one row of cells to the next (Figure 9-28). This phenomenon is most often a problem when the raster cell size is large relative to the buffer distance. A buffer distance of 100 m may be approximated when applied to a raster with a cell size of 30 m. A smaller cell size relative to the buffer distance results in less obvious "stair-stepping." The cell size should be small relative to the spatial accuracy of the data, and small relative to the buffer distance. If this rule is followed, then stair-stepping should not be a problem, because buffer sizes should be many times greater than the uncertainty inherent in the data.

Vector Buffers

Vector buffering may be applied to point, line, or area features, but regardless of input, buffering always produces an output set of area features (Figure 9-29). There are many variations in vector buffering. *Simple buffering*, also known as *fixed distance buffering*, is the most common form of vector buffering (Figure 9-29). Simple buffering identifies areas that are a fixed distance or greater from a set of input features. Simple buffering does not distinguish between regions that are close to one feature from those that are close to more than one feature. A location is either within a given distance from any one of a set of features, or farther away.

Simple buffering uses a uniform buffer distance for all features. A buffer distance of 100 meters specified for a roads layer may be applied to every road in the layer, irrespective of road size, shape, or location. In a similar manner, buffer distances for all points in a point layer will be uniform, and buffer distances for all area features will be fixed.

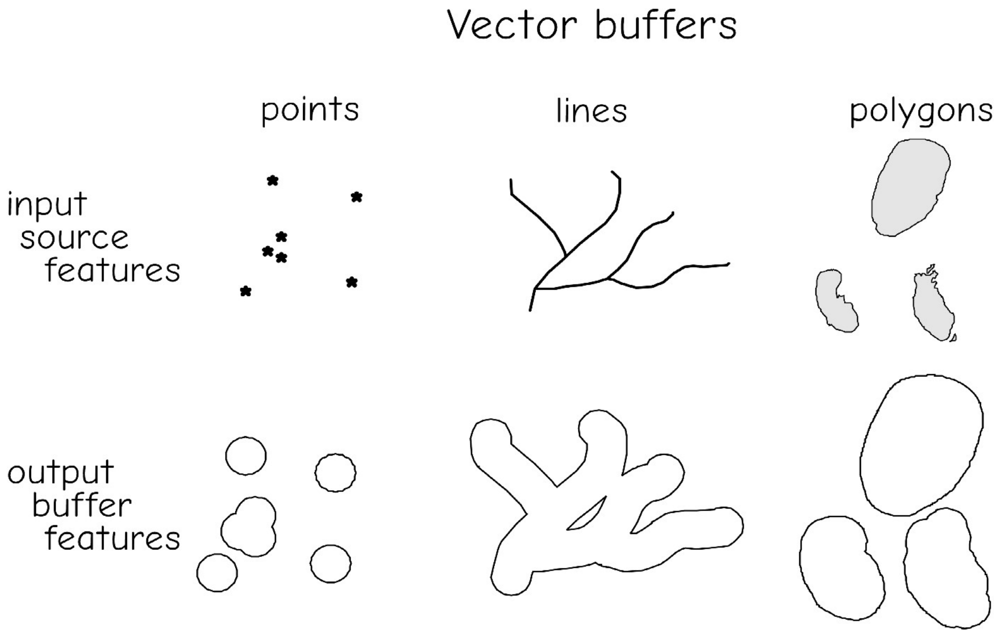


Figure 9-29: Vector buffers produced from point, line, or polygon input features. In all cases, the output is a set of polygon features.

Buffering on vector point data is based on the creation of circles around each point in a data set. The equation for a circle with an origin at x=0, y=0 is:

$$r = \sqrt{x^2 + y^2} \qquad (9.1)$$

where r is the buffer distance. The more general equation for a circle with a center at x_1, y_1, is:

$$r = \sqrt{(x - x_1)^2 + (y - y_1)^2} \qquad (9.2)$$

Equation (9.2) reduces to equation (9.1) at the origin, where x1 = 0, and y1 = 0. The general equation creates a circle centered on the coordinates x1, y1, with a buffer distance equal to the radius, r. Point buffers are created by applying this circle equation successively to each point feature in a data layer. The x and y coordinate locations of each point feature are used for x_1 and y_1, placing the point feature at the center of a circle.

Buffered circles may overlap, and in simple buffering, the circle boundaries that occur in overlap areas are removed. For example, areas within 10 km of hazardous waste sites may be identified by creating a buffer layer. We may have a data layer in which hazardous waste sites are represented as points (Figure 9-30a). A circle with a 10 km radius is drawn around each point. When two or more circles overlap, internal boundaries are dissolved, resulting in noncircular polygons (Figure 9-30b).

More complex buffering methods may be applied. These methods may identify buffer areas by the number of features within the given buffer distance, or apply variable buffer distances depending on the characteristics of the input features. We may be interested in areas that are near multiple hazardous waste sites. These areas near multiple hazardous sites may entail added risk and therefore require special monitoring or treatment. We may be mandated to identify all areas within a buffer

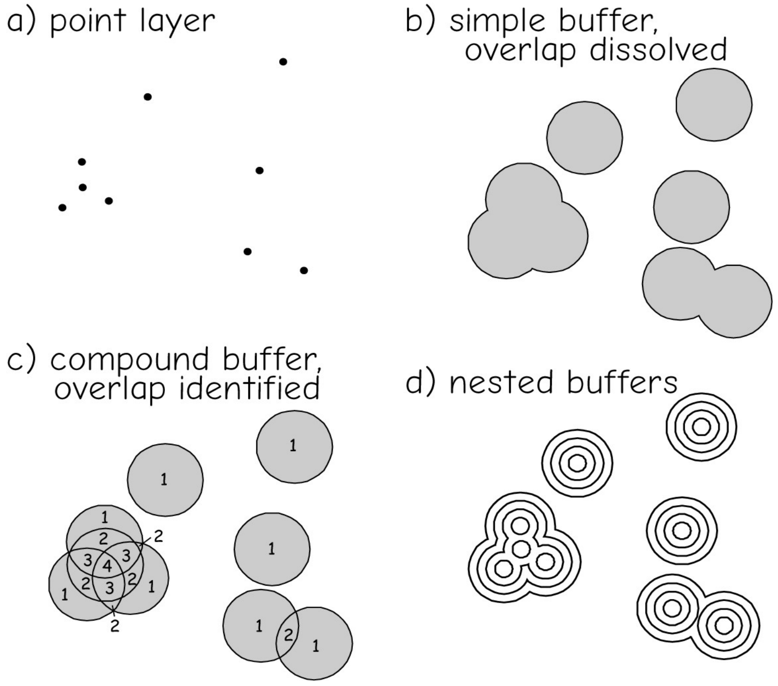


Figure 9-30: Various types of point buffers. Simple buffers dissolve areas near multiple features, more complex buffers do not. Multiring buffers provide distance-defined zones around each feature.

distance of a hazardous waste site, and the number of sites. In most applications, most of the dangerous areas will be close to one hazardous waste site, but some will be close to two, three, or more sites. The simple buffer, described above, will not provide the required information.

A buffering variant, referred to here as *compound buffering*, provides the needed information. Compound buffers maintain all overlapping boundaries (Figure 9-30c). All circles defined by the fixed radius buffer distance are generated. These circles are then intersected to form a planar graph. For each area, an attribute is created that records the number of features within the specified buffer distance.

Nested (or multiring) buffering is another common buffering variant (Figure 9-30d). We may require buffers at multiple distances. In our hazardous waste site example, suppose threshold levels have been established with various actions required for each threshold. Areas very close to hazardous waste sites require evacuation, intermediate distance require remediation, and areas farther away require monitoring. These zones may be defined by nested buffers.

Buffering on vector line and polygon data is also quite common. The formation of line buffers may be envisioned as a sequence of steps. First, circles are created that are centered at each node or vertex (Figure 9-31). Tangent lines are then generated. These lines are parallel to the input feature lines and tangent to the circles that are centered at each node or vertex. The tangent lines and circles are joined and interior circle segments dissolved.

Variable distance buffers (Figure 9-32) are another common variant of vector buffering. As indicated by the name, the buffer distance is variable, and may change among features. The buffer distance may increase in steps; for example, we may have one buffer distance for a given set of features, and a different buffer distance for the remaining features. In contrast, the buf-

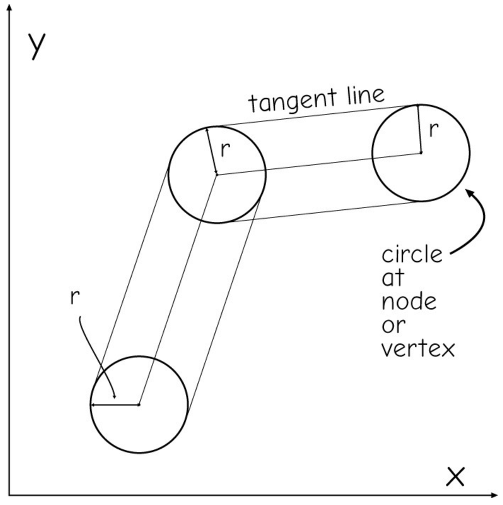


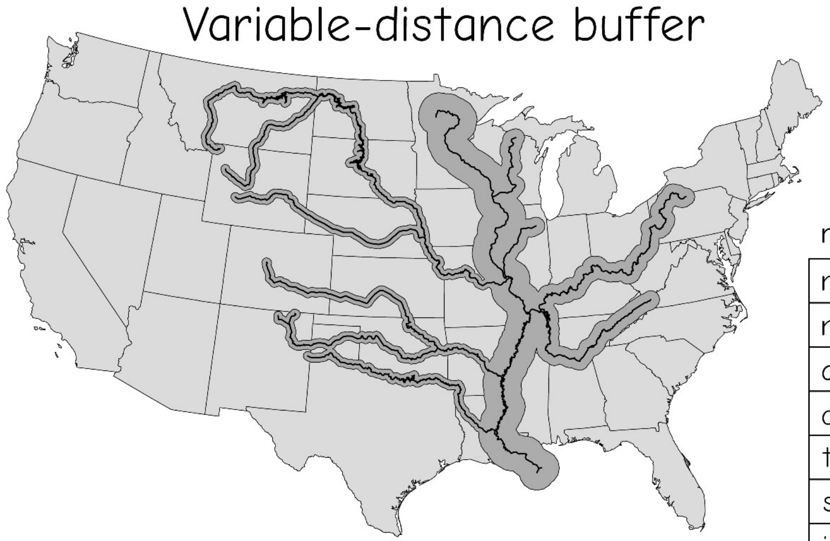
Figure 9-31: The creation of a line buffer at a fixed distance r.

fer distance may vary smoothly; for example, the buffer distance around a city may be a function of the population density in the city.

There are many instances for which we may require a variable distance buffer. We may wish to specify a larger buffer zone around large fuel storage facilities when compared to smaller fuel storage facilities. We often require more stringent protections further away from large rivers than for small rivers, and give large landfills a wider berth than small landfills.

Figure 9-32 illustrates the creation of buffers around a river network, where distance varies by river size. The increase in distance may be motivated by an increased likelihood of flooding downstream or an increased sensitivity to pollution. We may specify a buffer distance of 50 km for small rivers, 75 km for intermediate size rivers, and 100 km for large rivers. There are many other instances when variable distance buffers are required, for example, road noise, smoke stacks, or landfills.

The variable buffer distance is often specified by an attribute in the input data layer (Figure 9-32). A portion of the attribute table for the river data layer is shown. The attribute table contains the river name in river_identifier and the buffer distance is stored in buffdist. The attribute buffdist is accessed during buffer creation, and the size of the buffer adjusted automatically for each line segment. Note how the buffer size depends on the value in buffdist.



river_identifier	buffdist
mississippi	100
missouri	50
arkansas	50
ohio	75
tennessee	75
st. croix	75
illinois	75
wisconsin	75

Figure 9-32: An illustration of a variable distance buffer. A line buffer is shown with a variable buffer distance based on a river_identifier. A variable buffer distance, buffdist, is specified in a table and applied for each river segment.

Overlay

Overlay operations are powerful spatial analysis tools, and were an important driving force behind the development of GIS technologies. Overlays involve combining spatial and attribute data from two or more spatial data layers, and they are among the most common and powerful spatial data operations (Figure 9-33). Many problems require the overlay of thematically different data. For example, we may wish to know where there are inexpensive houses in good school districts, where whale feeding grounds overlap with proposed oil drilling areas, or the location of farm fields that are on highly erodible soils. In the latter example, a soils data layer may be used to identify highly erodible soils, and a current land use layer may be used to identify the locations of farm fields. The boundaries of erodible soils will not coincide with the boundaries of the farm fields in most instances, so these soils and land use data must somehow be combined. Overlay is the primary means of providing this combination.

An overlay operation requires that data layers use a common coordinate system. Overlay uses the coordinates that define each spatial feature to combine the data from the input data layers. The coordinates for any point on the Earth depend on the coordinate system used (Chapter 3). If the coordinate systems used in the various layers are not exactly the same, the features in the data layers will not align correctly.

Overlay may be viewed as the vertical stacking and merger of spatial data (Figure 9-33). Features in each data layer are set one "on top" another, and the points, lines, or area feature boundaries are merged into a single data layer. The attribute data are also combined so that the new data layer includes the information contained in each input data layer.

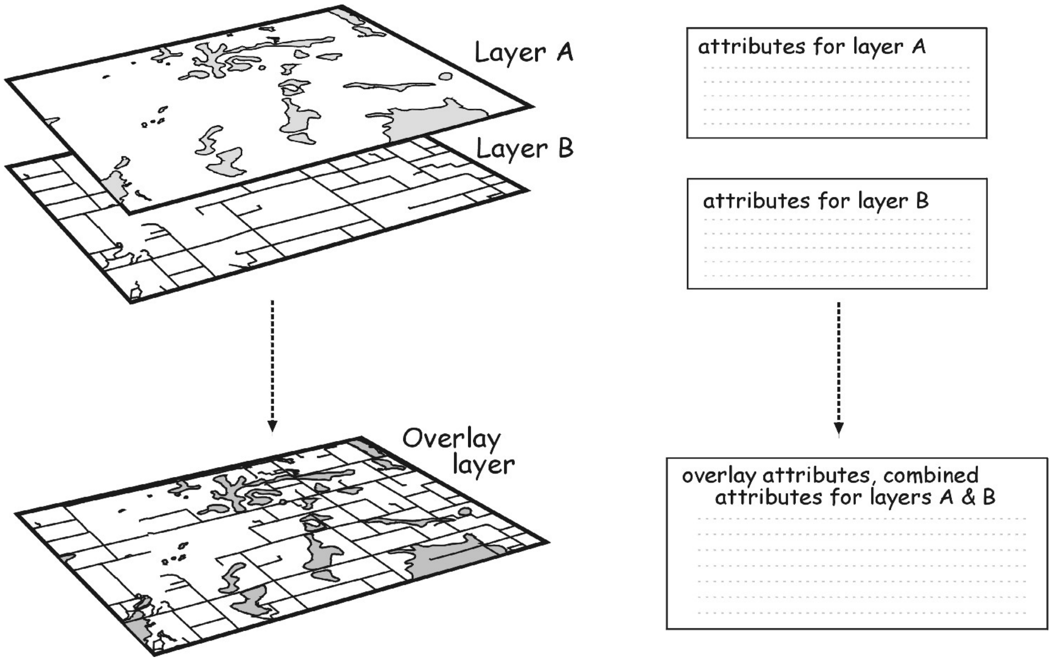


Figure 9-33: Spatial data overlay. Overlay combines both the coordinate information and the attribute information across different data layers.

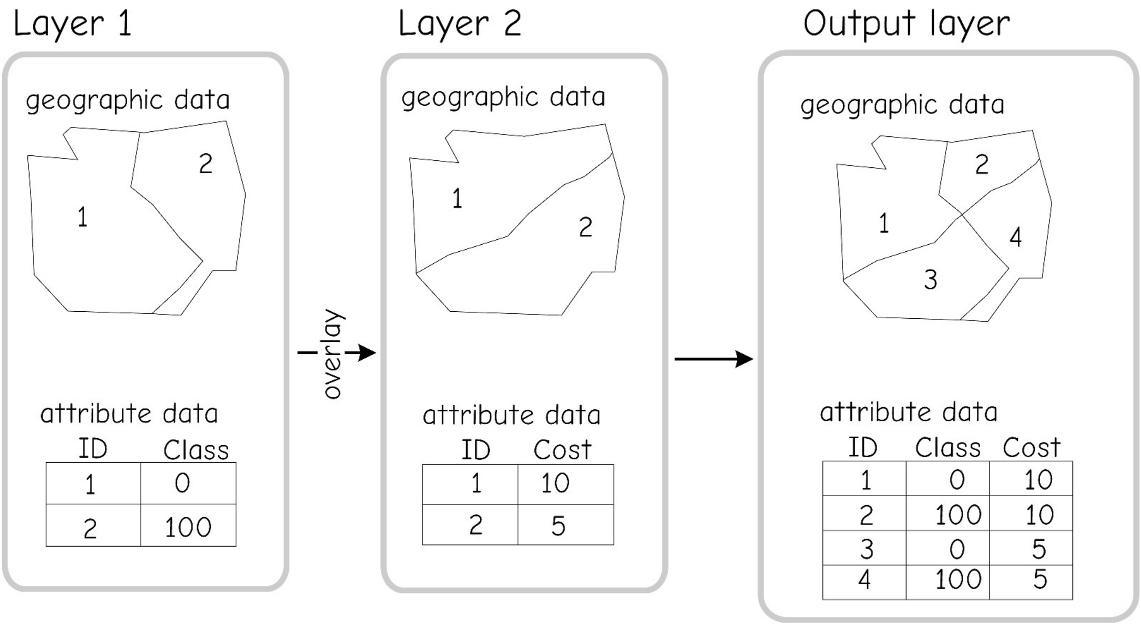
Vector Overlay

Overlay when using a vector data model involves combining the point, line, and polygon geometry and associated attribute data. This overlay creates new geometry. Overlay involves the merger of both the coordinate and attribute data from two vector layers into a new data layer. The coordinate merger may require the intersection and splitting of lines or areas and the creation of new features.

Figure 9-34 illustrates the overlay of two vector polygon data layers. This overlay requires the intersection of polygon boundaries to create new polygons. The overlay combines attribute data during polygon overlay. The data layer on the left is composed of two polygons. There are only two attributes for Layer 1, one an identifier (ID), and the other specifying values for a variable named class. The second input data layer, Layer 2, also contains two polygons, and two attributes, ID and cost. Note that the two tables have an attribute with the same name, ID. These two ID attributes serve the same function in their respective data layers, but they are not related. A value of 1 for the ID attribute in Layer 1 has nothing to do with the ID value of 1 in Layer 2. It simply identifies a unique combination of attributes in the output layer.

Vector overlay of these two polygon data layers results in four new polygons. Each new polygon contains the attribute information from the corresponding area in the input data layers. For example, note that the polygon in the output data layer with the ID of 1 has a class attribute with a value of 0 and a cost attribute with a value of 10. These values come from the values found in the corresponding input layers. The boundary for the polygon with an ID value of 1 in the output data layer is a composite of the boundaries found in the two input data layers. The same holds true for the other three polygons in the output data layer. These polygons are a composite of geographic and attribute data in the input data layers.

The topology of vector overlay output will likely be different from that of the input data layers. Vector overlay functions typically identify line intersections during



Layer 1 attribute data:

ID	Class
1	0
2	100

Layer 2 attribute data:

ID	Cost
1	10
2	5

Output layer attribute data:

ID	Class	Cost
1	0	10
2	100	10
3	0	5
4	100	5

Figure 9-34: An example of vector polygon overlay. In this example, output data contain a combination of the geographic (coordinate) data and the attribute data of the input data layers. New features may be created with topological relationships distinct from those found in the input data layers.

overlay. Intersecting lines are split and a node placed at the intersection point. Thus topology must be recreated if it is needed in further processing.

Any type of vector feature may be overlaid with any other type of vector feature, although some overlay operations rarely provide useful information and are performed infrequently. In theory, points may be overlaid on point, line, or polygon feature layers, lines on all three types, and polygons on all three types. Point-on-point or point-on-line overlay rarely results in intersecting features, and so they are rarely applied. Line-on-line overlay is sometimes required, for example, when we wish to identify the intersections of two networks such as road and railroads, but these also are rare occurrences. Overlays involving polygons are the most common by far.

Overlay output typically takes the lowest dimension of the inputs. This means point-in-polygon overlay results in point output, and line-in-polygon overlay results in line output. This avoids problems when multiple lower dimension features intersect with higher dimension features.

Figure 9-35 illustrates an instance where multiple points in one layer fall within a single polygon in an overlay layer. Output attribute data for a feature are a combination of the input data attributes. If polygons are output (Figure 9-35, right,

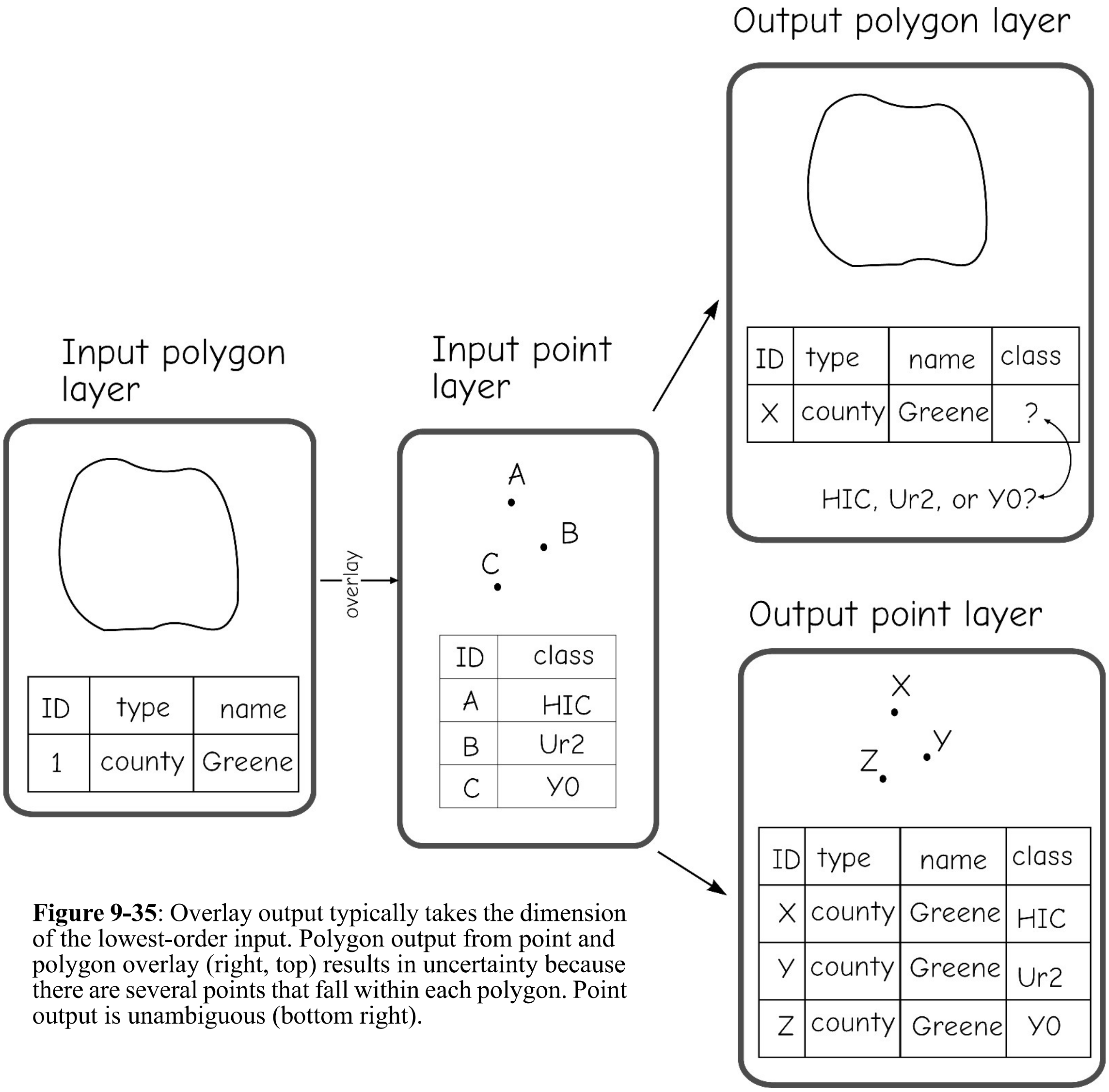


Figure 9-35: Overlay output typically takes the dimension of the lowest-order input. Polygon output from point and polygon overlay (right, top) results in uncertainty because there are several points that fall within each polygon. Point output is unambiguous (bottom right).

top), there is ambiguity regarding which point attribute data to record. Each point feature has a value for an attribute named class. It is not clear which value should be recorded in the output polygon, the class value from point A, point B, or point C. When a point layer is output (Figure 9-35, right, bottom), there is no ambiguity. Each output point feature contains the original point attribute information, plus the input polygon feature attributes.

One method for creating polygon output from point-in-polygon overlay involves recording the attributes for one point selected arbitrarily from the points that fall within a polygon. This is usually not satisfactory because important information may be lost. An alternative involves adding columns to the output polygon to preserve multiple points per polygon. However, this would still result in some ambiguity, such as, what should be the order of duplicate attributes? It may also add a substantial number of sparsely used items, thus increasing file size inefficiently. Forcing the lower order output during overlay avoids these problems, as shown in the lower right of Figure 9-35.

Note that the number of attributes in the output layer increases after each overlay. This is illustrated in Figure 9-35, with the combination of a point and polygon layer in an overlay. The output point attribute table shown in the lower right portion of the figure contains four items. This output attribute table is a composite of the input attribute tables.

Large attribute tables may result if overlay operations are used to combine many data layers. When the output from an overlay process is in turn used as an input for a subsequent overlay, the number of attributes in the next output layer will usually increase. As the number of attributes grows, tables may become unwieldy, and we often delete redundant attributes.

Overlays that include a polygon layer are most common. We are often interested in the combination of polygon features with other polygons, or in finding the coincidence of point or line features with polygons. What counties include hazardous waste sites? Which neighborhoods does one pass through on E Street? Where are there shallow aquifers below cornfields? All these examples involve the overlay of area features, either with other area features, or with point or line features.

Clip, Intersect, and Union: Special Cases of Overlay

There are three common ways overlay operations are applied: as a *clip*, an *intersection*, or a *union*.

The basic layer-on-layer combination is the same for all three. They differ in the geographic extent for which vector data are recorded, and in how data from the attribute layers are combined. Intersection and union are derived from general set theory operations. The intersection operation may be considered in some ways to be a spatial AND, while the union operation is related to a spatial OR. The clip operation may be considered a combination of an intersection and an elimination. All three are common and supported in some manner as stand-alone functions by most GIS software packages.

A *clip* may be considered a "cookie-cutter" overlay (Figure 9-36). A bounding polygon layer is used to define the areas for which features will be output. This bounding polygon layer defines the clipping region. Point, line, or polygon data in a second layer are "clipped" with the bounding layer. In most versions of the clip function, the attributes for the clipping layer are not included in the output data layer. A clip is most often used when sub-setting data geographically, to reduce data volumes when working on a small area included in larger data layers. A city manager may only wish the set of streets within their city boundaries, clipped from a statewide roads layer.

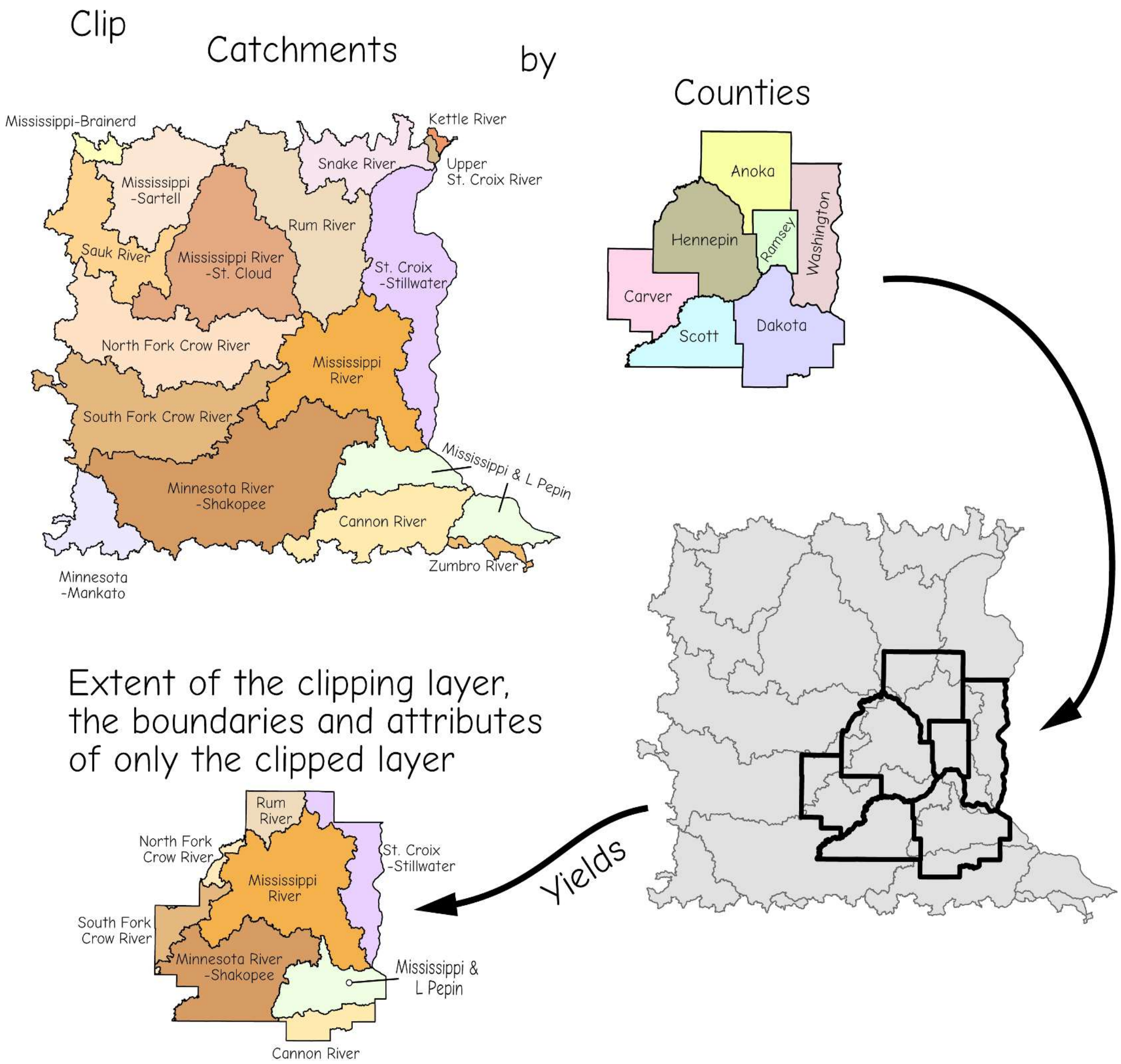


Figure 9-36: A clip is a common variation of overlay operations. The clip preserves information only from the clipped (or target) data layer and only for the area of the clipping (or bounding layer). The attribute table of the output layer typically contains all the attributes of the clipped layer (here, Catchments), and none from the clipping layer (Counties).

In our example shown in Figure 9-36, the bounding or clipping data layer consists of seven county polygons, and the target or clipped data layer contains many small catchment boundaries. The presence of polygon attributes in the bounding layer is indicated by the different shades for the different county polygons. The output from the clip consists of those portions of catchments within the clip layer boundary. Note that the clip layer boundaries, here counties, are not included in the output data layer. Also note that only the attributes for the clipped catchment layer are output.

Users should be certain that transferred variables still have valid values after a clip. If an area field is included in the input layer, the value may be wrong if it is not recalculated. Other area-based values may also be in error, e.g., a polygon density or counts of included features. Since software defaults vary, the behavior of the specific software tool should be identified.

An *intersection* is another multi-layer combination, and may be defined as an overlay that fuses data from both layers, but only for the area where both layers contain data (Figure 9-37). Features boundaries from both data layers are combined. Both layers serve as data and as bounding layers, so that any parts of polygons that are in one layer but not another are clipped and discarded.

The same caution on relevancy and recalculation of combined, area-based variables applies to the output from intersection operations as applied to clip operations. Since new geographies with differing areas are created, attributes transferred from the component features may need new interpretations. Most implementations simply copy the values from the input features to the coincident output features. For example, a county area for each polygon will come from the input county polygon. While this was valid for each input county in Figure 9-37, it will be replicated for each polygon in the output layer, and should be interpreted as the area for the contributing county, not the area from that county in the (usually smaller) output polygon. Summing across polygons with a

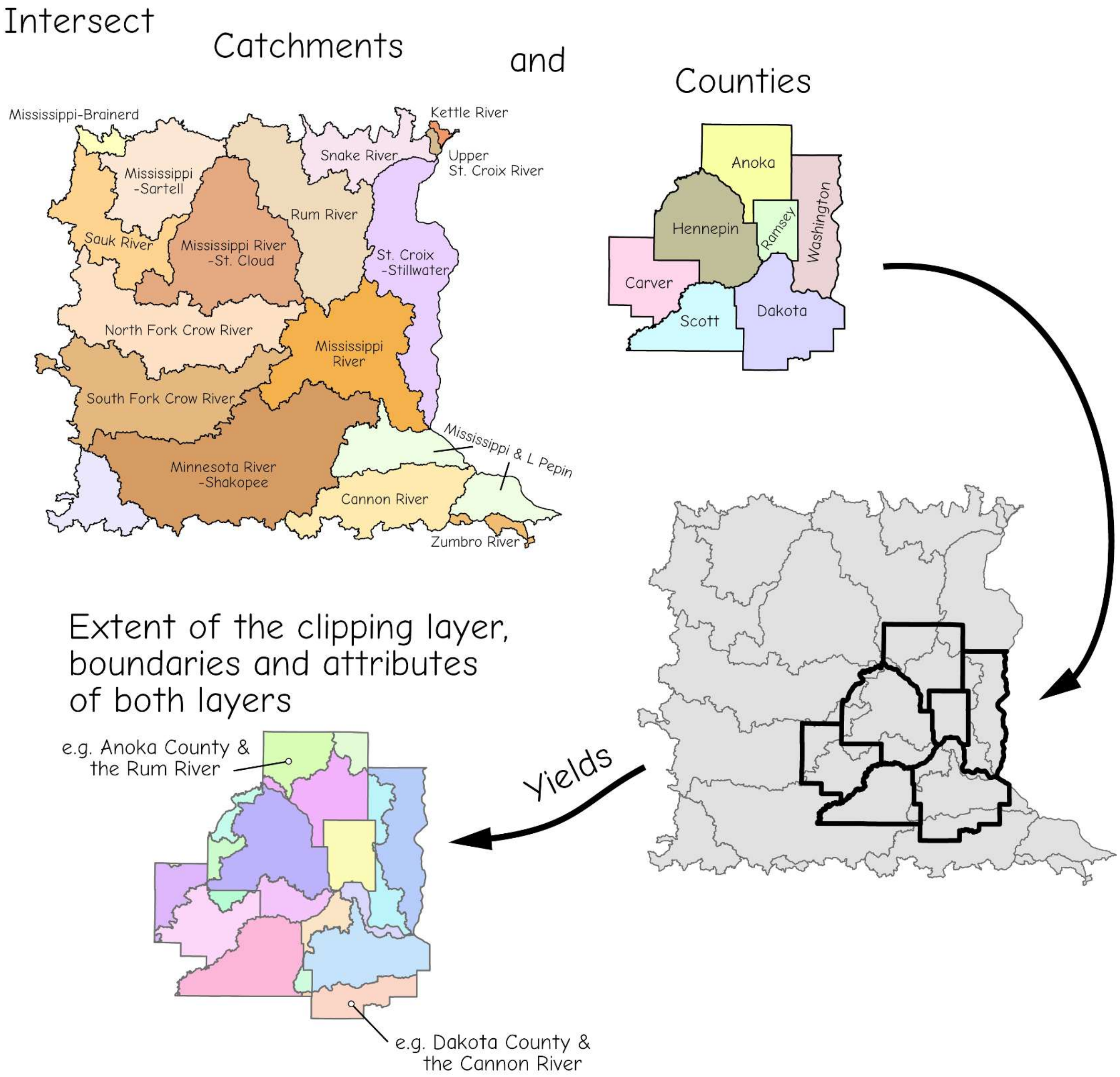


Figure 9-37: An intersection is another common overlay operation. Both the boundaries and data are combined in the output, but only for the areas that are contained in the clipping layer, Counties in this example. The particular software tool typically requires you explicitly identify the clipping layer.

given county value will usually not give an accurate county area. Each output variable should be scrutinized and the value origin and contents clearly understood.

A *union* is another kind of overlay. It retains all data from both the bounding and data layers (Figure 9-38). No geographic data are discarded in the union operation, and corresponding attribute data are saved for all regions. New polygons are formed by the combination of coordinate data from each data layer.

Note that there are often many null or empty attribute values in union output layers. Data in non-overlapping areas are absent and so cannot be assigned, e.g., outside the county layer bounds but within the watershed layer bounds in Figure 9-38. The presence of null values may alter subsequent operations.

Many software packages support additional variants of overlay operations. Some support an *Erase* or similarly-named function, which is the complement to the clip function. In an Erase function, areas covered by the input layer are "cut out" or

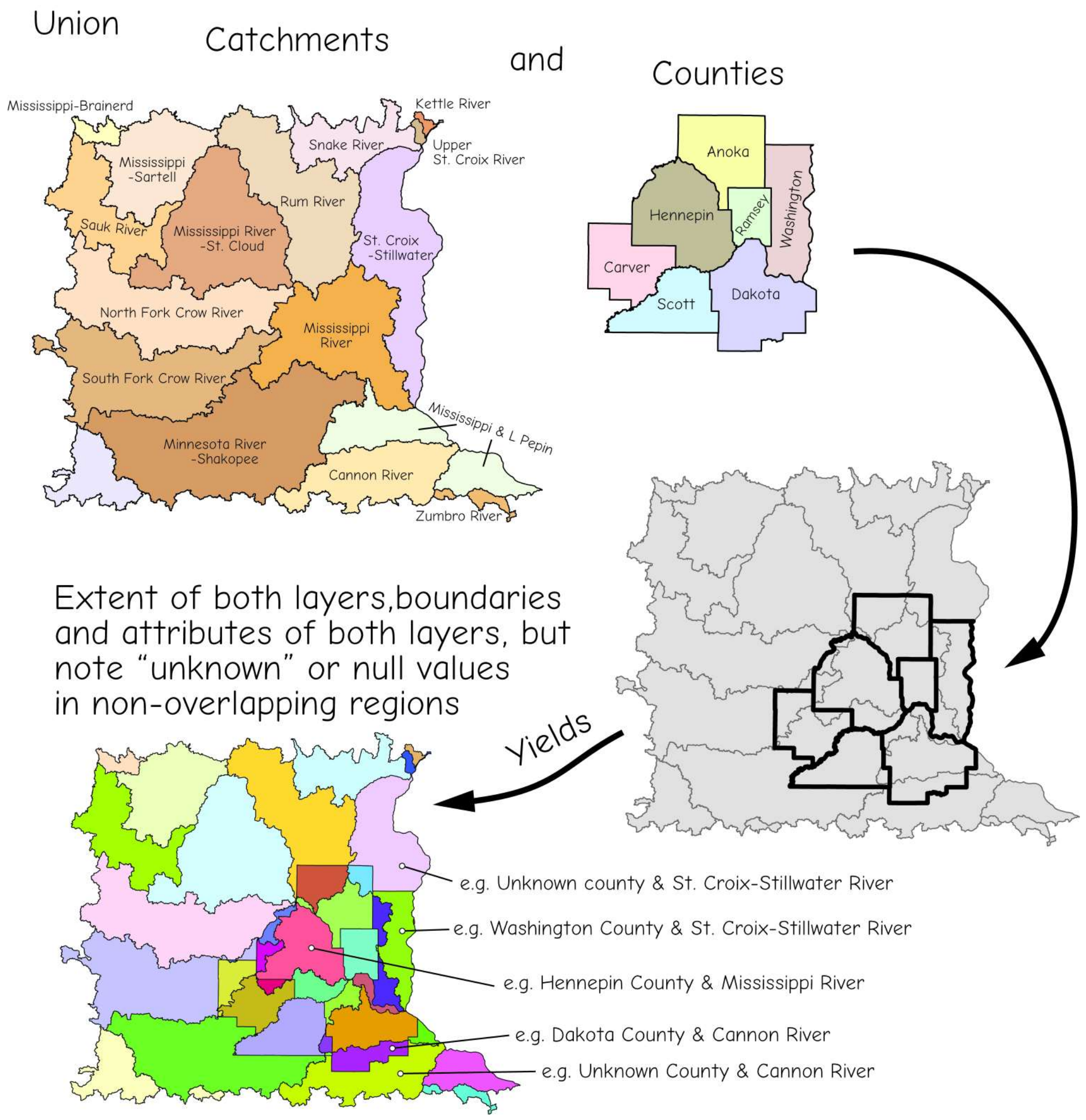


Figure 9-38: A union operation. Data from two layers are combined, including any areas that are contained by one but not both data layers. No areas nor attribute data are discarded, or "clipped," as in the intersection and clip operations. Blank or null values are typically assigned in the areas contained in one but not the other layer, e.g., in the example where the county is listed as "Unknown."

erased from the bounding layer (Figure 9-39). Erasures may cut existing polygons apart or, where there are coincident lines in the two data layers, may preserve the edge of existing polygons. In some versions, there is a tolerance distance that allows for lines that aren't exactly coincident in different layers, but are meant to be, to be represented only once in the output. This tolerance distance effectively serves as a snapping distance, and moves the vertices in one layer on a nearly coincident edge to match vertices in the other layer. As with snapping during digitization or other overlays, this may help reduce incorrect or unwanted geometries.

The *Erase* function is particularly useful when updating a portion of a data layer, in that old, out of date, poorer quality, or otherwise inferior data may be clipped out of a section and newer data substituted. Erase is also often useful in spatial analyses that include criteria specifying areas that are greater than some distance from a set of

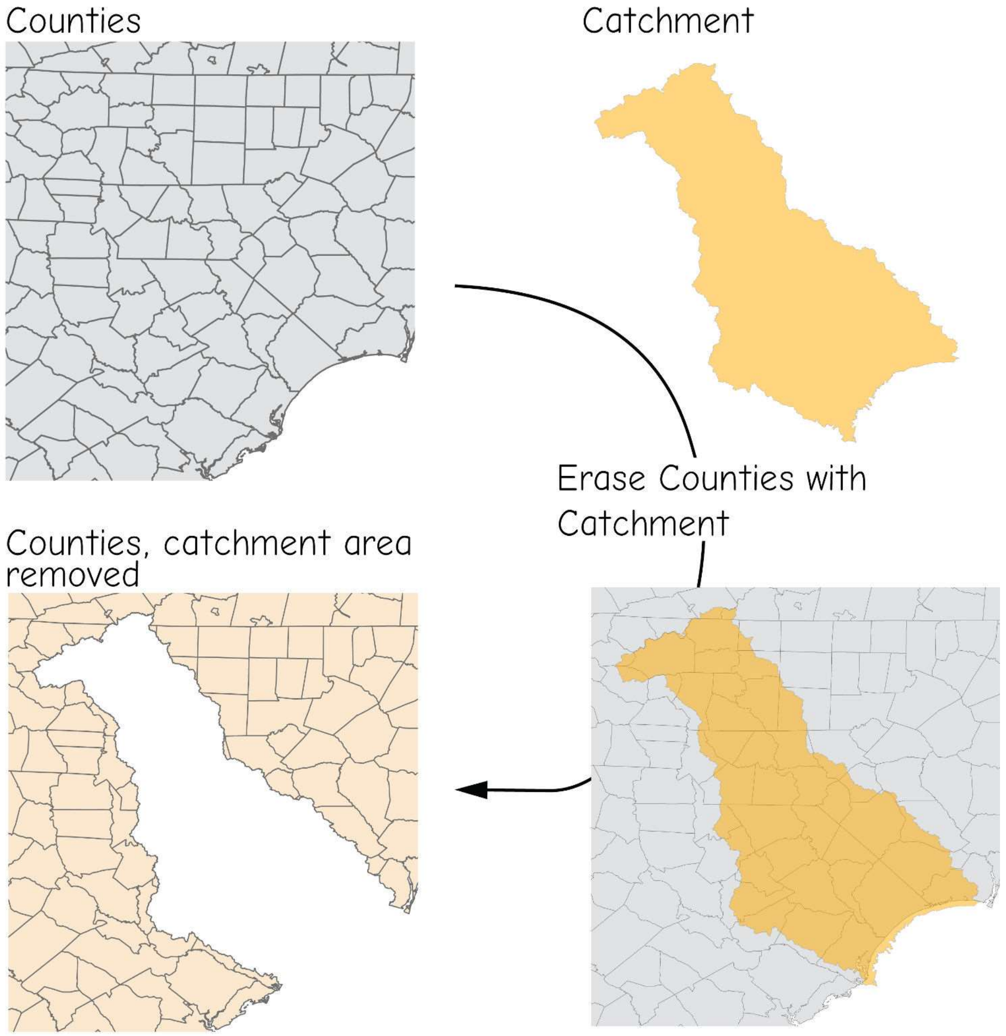


Figure 9-39: An example of an Erase operation. Features in an input layer are "erased" based on the outer boundary of an erasing polygon. This operation is often used in editing or cartographic models that specify area removal based on a buffered layer.

features. A buffer function identifies areas that are less than the target distance, and these may then be removed from consideration using an erase operation.

There are other variants on unions or intersections. Most of these specialized overlay operations may be created from the application of union or overlay operations in combination with selection operations.

Vector overlay is often a time-consuming computational process, due to the large number of lines that must be compared. A vector overlay typically requires repeated tests of line intersection, a relatively simple set of calculations, but there is often a large number of line segments in a data set. Each line segment must be checked against every other line segment, requiring perhaps billions of tests for line intersection.

A Problem in Vector Overlay

Polygon overlays often suffer when there are common features that are represented in both input data layers. We define a common feature as a different representation of the same phenomenon. Figure 9-40 illustrates this problem. A county boundary may coincide with a state boundary. However, different versions of the state and county boundaries may be created independently from two adjacent states, using different source materials, at different times, and using different systems. Thus, these two representations may differ even though they identify the same boundary on the earth surface.

In most data layers, the differences will be quite small, and will not be visible except at very large display scales, for example, when the on-screen zoom is quite

Figure 9-40: Sliver polygons may occur when two representations of a feature are combined. A common boundary between two features has been derived from different sources. The representations differ slightly. This results in small gaps and "sliver" polygons along the margin between these two layers.

high. The differences are shown in the larger-scale inset in Figure 9-40. When the county and state data layers are overlaid, many small polygons are formed along the boundary. These polygons are quite small, but they are often quite numerous.

These "sliver" polygons cause problems because there is an entry in the attribute table for each polygon. One-half or more of the polygons in the output data layer may be these slivers. Slivers are a burden because they take up space in the attribute table but are not of any interest or use. Analyses of large data sets are hindered because all selections, sorts, or other operations must treat all polygons, including the slivers. Processing times often increase exponentially with the number of polygons.

There are several methods to reduce the occurrence of these slivers. One identifies all common boundaries across different layers. The boundary with the highest coordinate accuracy is substituted into all other data layers, replacing the less accurate representations. This involves considerable editing, and most common when developing new data layers.

Another method involves manually identifying and removing slivers. Small polygons may be selected, or polygons with two bounding arcs, common for sliver polygons. Bounding lines may then be adjusted or removed. However, manual removal is not practical for many data sets due to the high number of sliver polygons.

A third method for sliver reduction involves defining a snap distance during overlay. Much as with a snap distance used during data development (described in Chapter 4), this forces nodes or lines to be coincident if they are within a specified proximity during overlay. As with data entry, this snap distance should be small relative to the spatial accuracy of the input layers and the required accuracy of the output data layers. If the two representations of a line are within the snap distance then there will be no sliver polygons. In practice, not all sliver polygons are removed, but their numbers are substantially reduced, thereby reducing the time spent on manual editing.

Automatic sliver detection and removal should be applied carefully, as they may delete valuable data. Only small slivers should be removed, with small defined as smaller than an area, length, or width that is worth tracking. This distance may be set by the accuracy of the data collection, or by the requirements of the analysis. If polygon edge locations are only digitized to within one meter of their true position, it makes little sense to maintain polygons that are less than a meter in any dimension. However, if slivers are removed that are substantially wider and longer than a meter, some valuable information may be lost.

Raster Overlay

Raster overlay involves the cell-by-cell combination of two or more data layers. Data from one layer in one cell location correspond to a cell in another data layer. The cell values are combined in some manner and an output value assigned to a corresponding cell in an output layer.

Raster overlay is typically applied to nominal or ordinal data. A number or character stored in each raster cell represents a nominal or ordinal category. Each cell value corresponds to a category for a raster variable. This is illustrated in the input data sets shown at the left and center of Figure 9-41. Input Layer A represents soils data. Each raster cell value corresponds to a specific soil value. In a similar manner, input Layer B records land use, with values 1, 2, and 3 corresponding to particular land uses. These data may be combined to create areas fusing the two input layers – cells with values for both soil type and land use.

There are as many potential output categories as there are possible combinations of input layer values. In Figure 9-41 there are two soil types in Layer A, and three land use types in Layer B. There may be 3 x 2, or 6 different combinations in the output layer. Not all combinations will necessarily occur in the overlay, as shown in Figure 9-41. In this example only four of the six overlay combinations occur. Unique identifiers must be generated for each observed combination, and placed in the appropriate cell of the output raster layer.

The number of possible combinations is important to note because it may change the number of binary digits or bytes required to represent the output raster data layer. A raster cell typically contains a number or character, and may be a one-byte integer, a two-byte integer, or some other size. Raster data sets typically use the smallest required data size. As discussed in Chapter 2, one unsigned byte may store up to 256 different values. Raster overlay may result in an output data layer that requires a higher number of bytes per cell. Consider the overlay between two raster data layers, one layer that contains 20 different nominal classes, and a second layer with 27 different nominal classes. There is a total of 20 times 27, or 540, possible output combinations. If more than 256 combinations occur, the output data will require more than one byte for each cell. Typically two bytes will

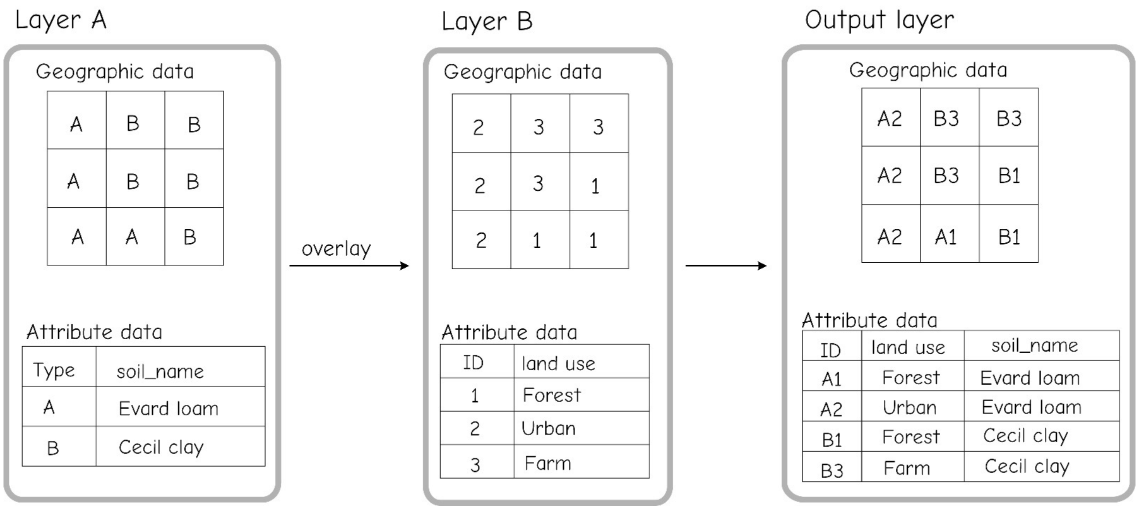


Figure 9-41: Cell-by-cell combination in raster overlay. Two input layers are combined in raster overlay. Nominal variables for corresponding cells are joined, creating a new output layer. In this example a soils layer (Layer A) is combined with a land use layer (Layer B) to create a composite Output layer.

be used. This causes a doubling in the output file size. Two bytes will hold more than 65,500 unique combinations; if more categories are required, then four bytes per cell are often used.

Raster overlay requires that the input raster systems be compatible. This typically means they should have the same cell dimension and coordinate system, including the same origin for x and y coordinates. If the cell sizes differ, there will likely be cells in one layer that match parts of several cells in the second input layer (Figure 9-42). This may result in ambiguity when defining the input attribute value. Overlay may work if the cells are integer multiples with the same origin, for example, the boundaries of a 1 by 1 meter raster layer may be set to coincide with a 3 by 3 raster layer; however this rarely happens. Data are normally converted to compatible raster layers before overlay. This is most often done using a resampling, as described in Chapter 4. In our example, we might choose to resample Layer 2 to match Layer 1 in cell size and orientation. Values for cells in Layer 2 would be combined through a nearest neighbor, bilinear interpolation, cubic convolution, or some other resampling formula to create a new layer based on Layer 2 but compatible with Layer 1.

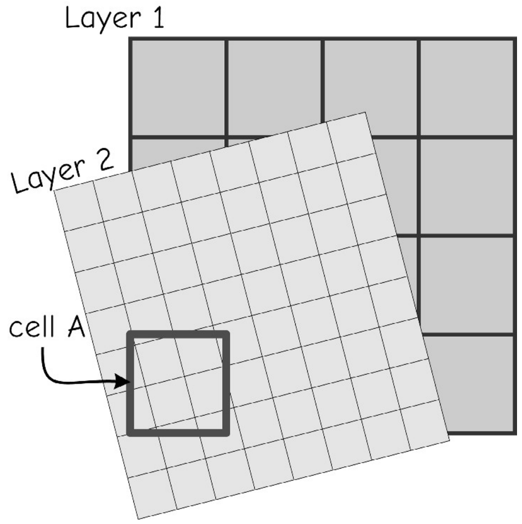
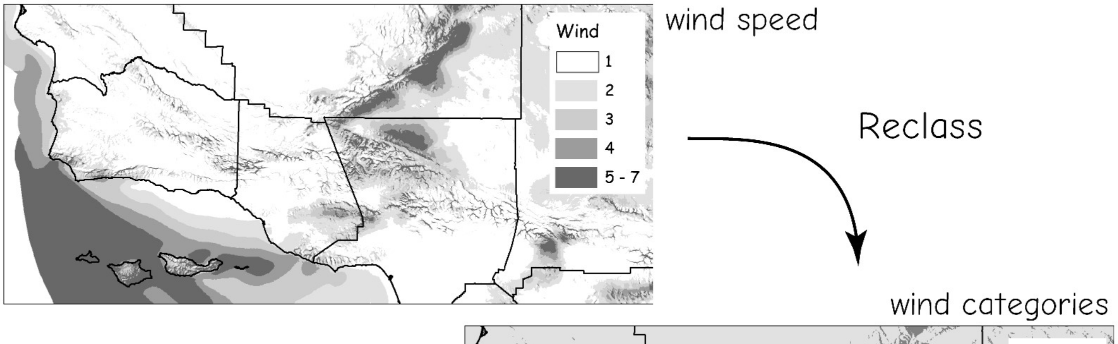


Figure 9-42: Overlaid raster layers should be compatible to ensure unambiguous overlay. In the overlay depicted here it is not clear which cells from Layer 2 should be combined with cell A in Layer 1.

An Example Spatial Analysis

Figure 9-43 and the following figures briefly illustrate an application of basic spatial analysis. We seek to identify suitable areas for wind farms, based on two criteria: areas with high average wind speeds, and areas with a low population density. High average winds are preferred because the energy produced at a site increases with wind speed. Low population densities are preferred because land is less expensive and there are fewer neighbors to bother. This simple example does not include obvious additional factors, such as the distance to power lines, avoiding protected lands, or the difficulties of building offshore vs. onshore, but it does illustrate how data may be combined in a set of simple spatial functions to answer a question. This analysis requires wind data of appropriate accuracy, spatial extent, and appropriately summarized. For example, I might wish to base my analysis on average daily wind speed, or maximum hourly wind speed for a day, or maximum daily wind speed. If these data do not exist, I would need to either change the problem formulation, my analysis methods, or develop the data from existing sources. For example, if a gridded data set does not exist, but there are point observations from a network of weather stations, I might use interpolation or other methods (described in Chapter 12) to estimate wind speed across my study region. These considerations highlight an important early step in spatial analysis: we must assess the available data, and determine if it is appropriate for our intended analysis. If not, we must create the required data or change our analysis.

For this example, wind data were obtained from the U.S. Department of Energy, and population data from the U.S. Census Bureau. Wind data were reclassed to those values (4 or greater) that provided suitable potential energy (Figure 9-43). We might represent this graphically as shown in Figure 9-44, where the input and output layers are shown as boxes, and the spatial operation noted inside an ellipse. Arrows show the direction or flow of the analysis. The categories used in the reclassification should be based on prior knowledge; here, we might know that wind levels below the category 4 are unsuitable for the wind turbines to be used. In general, thresholds or



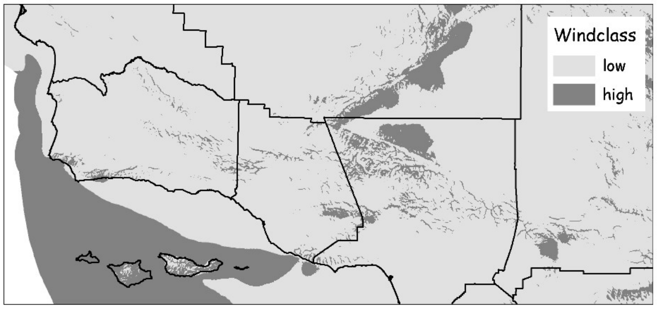


Figure 9-43:Areas with suitable average wind speeds to be considered further. A threshold is determined and used in a reclassification operation to identify acceptable areas.

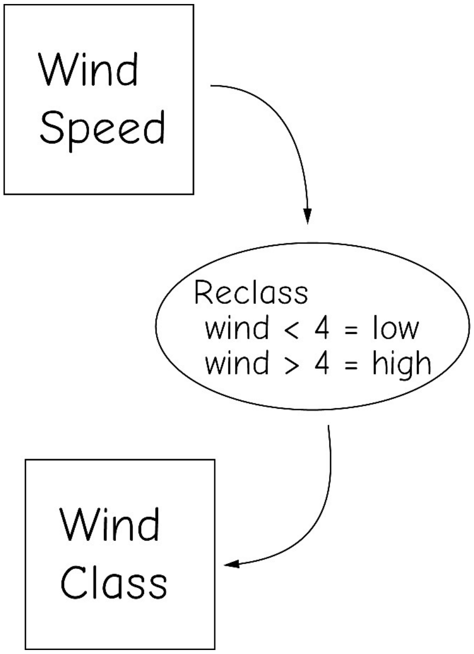
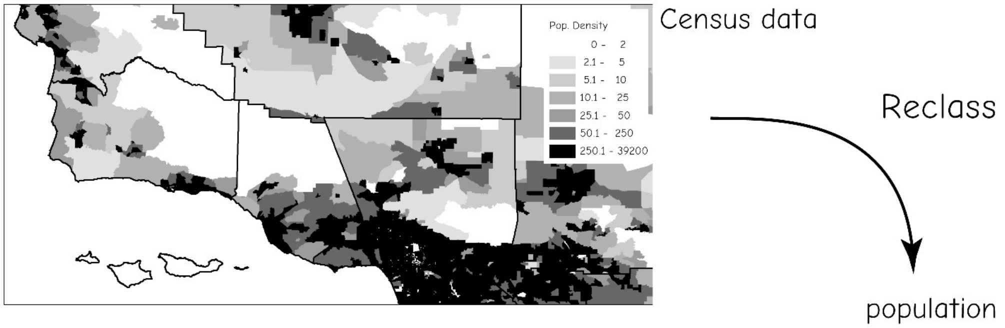


Figure 9-44: A graphic representation of the spatial operation to create a wind class layer from wind speed data.

selection values depend on the problem under consideration, and should be well-justified by additional background information.

Selection of areas with a suitably sparse population is shown in Figure 9-45. Here, data are placed into two categories, sparse and dense. As with the wind classification, the classes should be based on some external information, e.g., land prices drop or the likelihood of irate neighbors drops below a specific population density threshold. It also assumes the input census data layer provides polygons with population density suitably calculated, for example, in persons per square mile. If not, this calculation will have to be performed prior to the reclassification.



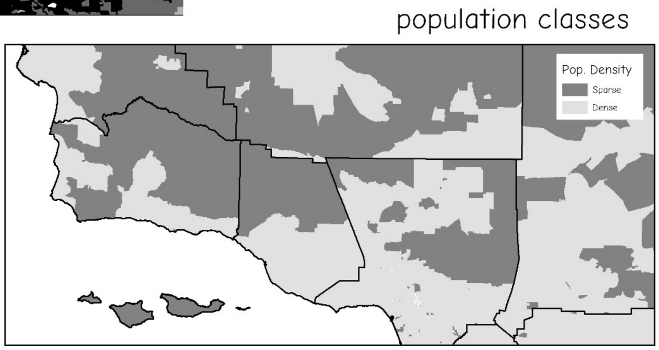


Figure 9-45: Areas with sufficiently sparse populations were identified by reclassifying U.S. census data.

Here the threshold for density is set at 10 persons per square mile, and polygons reclassed accordingly, creating a new data layer. A graphic representation of the spatial operation is shown in Figure 9-46.

Reclassed layers were then combined in an overlay operation, and selected to identify areas that have both low population densities and high wind speeds (Figure 9-47). In practice, this will involve several steps in a spatial operation, with separate overlay, selection, and reclassification steps. These steps are abbreviated in Figure 9-47, showing just one general operation.They are more fully sketched in Figure 9-48

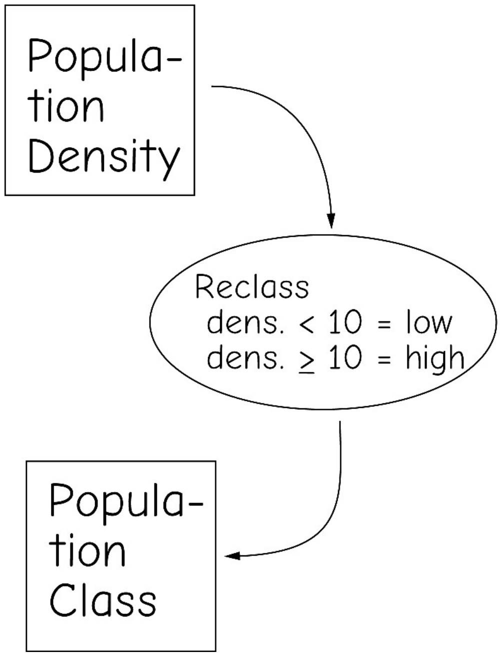


Figure 9-46: Diagram representing the population reclassification.

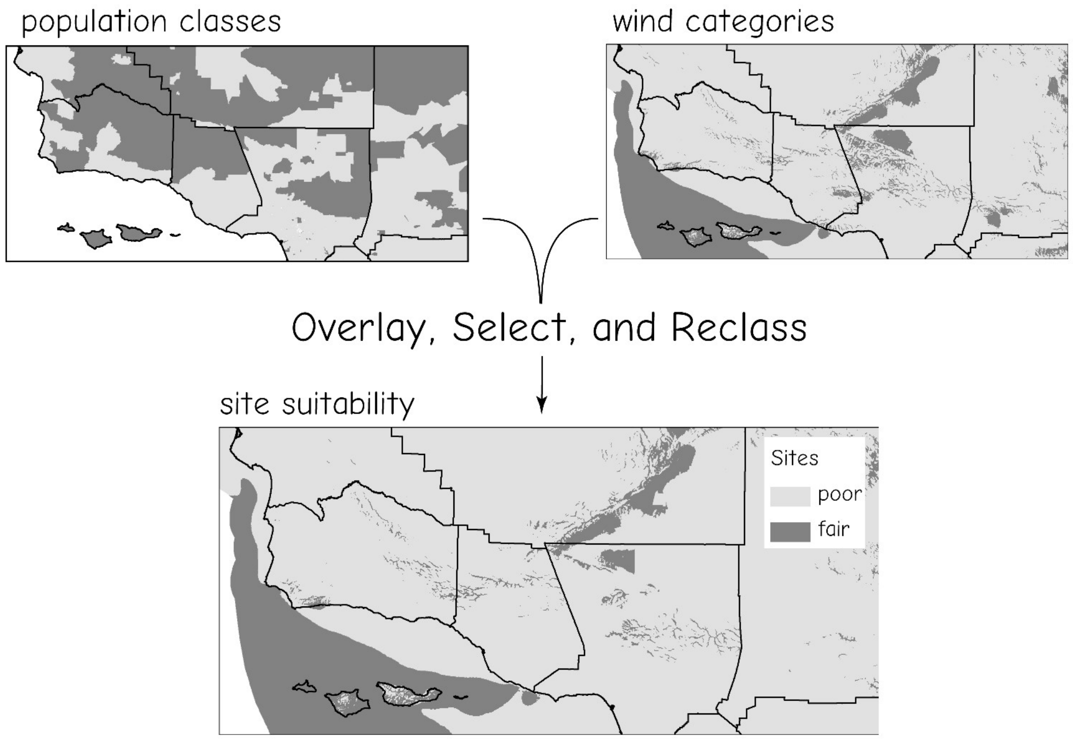


Figure 9-47: These intermediate layers are combined in an overlay operation, and then areas selected based on criteria for each resultant geographic unit.

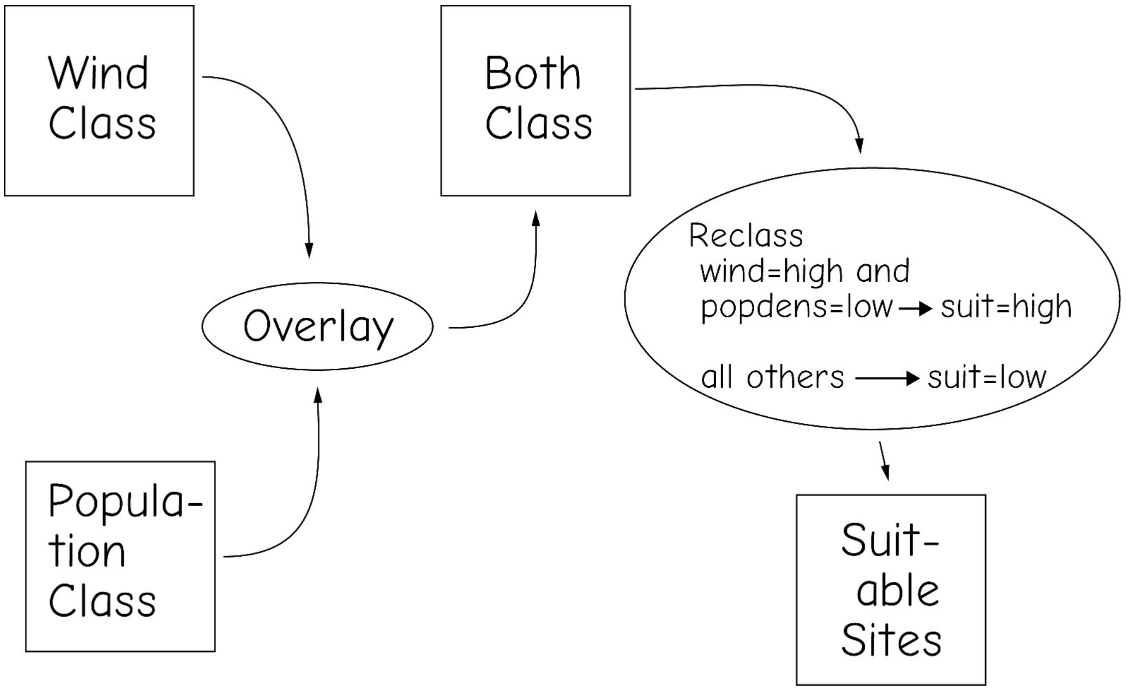


Figure 9-48: A graphic representation of the final steps in the example spatial analysis. This graphic expands on Figure 9-47 to explicitly show the overlay and reclassification operations, and the intermediate layer required conceptually, and when using most currently available software.

Network Analysis

Networks are common in our lives. Roads, powerlines, telephone and television cables, and water distribution systems are all examples of networks we utilize many times each day (Figure 9-49). As networks are crucial to civilization, they need to be effectively managed. These networks also represent substantial investments, and their management merits considerable attention. Spatial analysis tools have been developed to help us use and maintain networks.

A *network* may be defined as a set of connected features, often termed *centers*. These features may be centers of demand, centers of supply, or both (Figure 9-50). Centers are connected to at least one and possibly many *network links*. Links interconnect and provide paths between centers. Traveling from one center to another often requires traversing many separate links.

Network analyses, also known as network models, are used to represent and analyze the cost, time, delivery, and accumulation of resources along links and between the connected centers. Resources flow to and from the centers through the networks. In addition, resources may be generated or absorbed by the links themselves.

The links that form the networks may have attributes that affect the flow. For example, there may be links that slow or speed up the flow of resources, or a link may allow resources to flow in only one direction. Link attributes are used to model flow characteristics of the real network; for example, travel on some roads is slower than others, or cars may legally move in only one direction on a one-way street.

The concept of a *transit cost* is key to many network analysis problems. A transit cost reflects the price one pays to move a resource through a segment of the network. Transit costs are typically measured in time, distance, or monetary units; for example, it costs 10 seconds to travel through a link. Costs may be constant such that it always

Figure 9-49: Networks in a GIS are used to represent roads, pipelines, power transmission grids, rivers, and other connected systems through which important resources flow.

takes 10 seconds to traverse the link regardless of direction or time of day. Alternatively, costs may vary by time of day or direction, so it may take 15 seconds to traverse an arc during morning and evening rush hours, but 10 seconds otherwise, or it may take twice as long to travel north to south than to travel south to north.

We will discuss three types of problems that are commonly analyzed using networks: route selection, resource and territory allocation, and traffic modeling. There are many types of networking problems; however, these three are among the most common and provide an indication of the methods and breadth of network analyses.

Route selection involves identifying a "best" route based on a specified set of criteria. Route selection is often applied to find the least costly route that visits a number of centers. Two or more centers are identified within a network, including starting and ending centers. These centers must all be visited by traversing the network. There are usually a very large number of alternative routes, or pathways, that may be used to visit all centers. The best route is selected based on some criteria, usually the shortest, quickest, or least costly route. Further restrictions may be placed on the route; for example, the order in which centers are visited may be specified.

Route selection may be used to improve the movement of public transportation through a network. School buses are often routed using network analyses. Each bus must start and finish at a school (a center) and pick up children at a number of stops (also centers). The shortest path or time route may be specified. Alternate routes are analyzed and the "best" route selected.

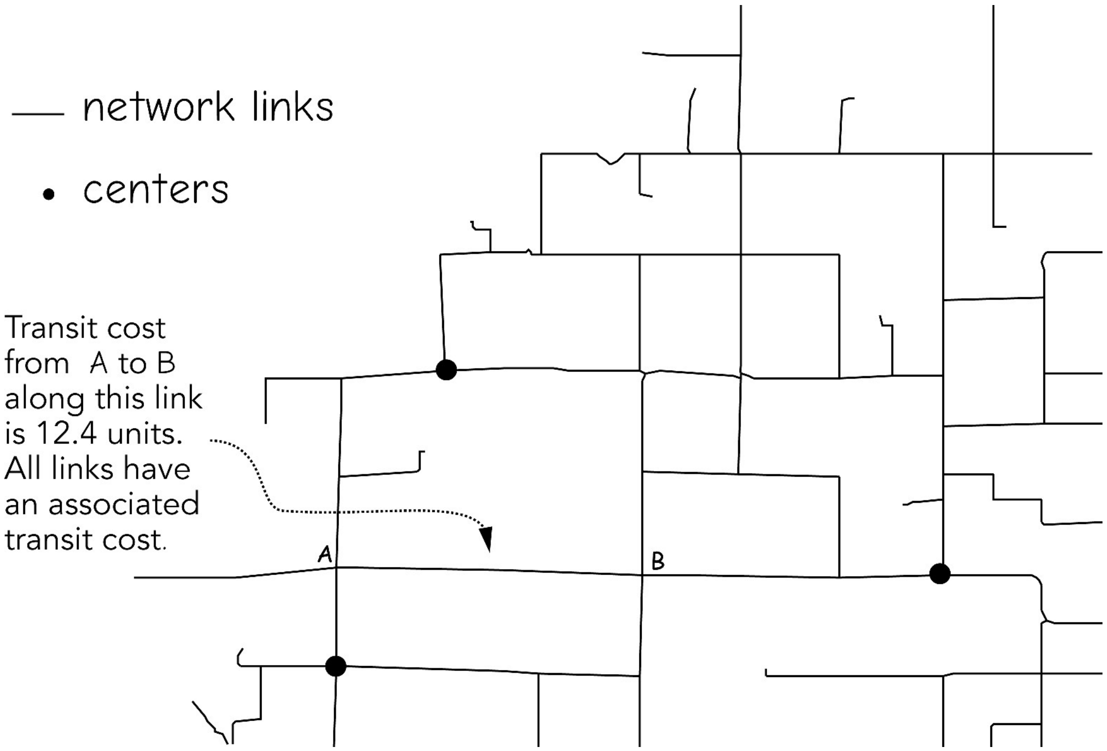


Figure 9-50: Basic network elements. Centers are connected by a set of links. Costs may be associated with traversing the links. Network analysis typically involves moving resources or demands among centers.

Selection of the best route involves an algorithm that recursively follows a least-cost set of arcs, beginning at the current node. A set of interconnected network links is identified, as well as start and destination centers (Figure 9-51). The route from start to destination locations is typically built iteratively. One route finding algorithm adds the least-cost link at each step. Multiple paths are tested until a path connects the start and destination centers.

This simple method begins at the start center. Paths are extended by adding the link that gives the lowest total cost for all paths currently pursued. The initial set of candidate links consists of all those connecting to the starting point. The lowest cost link is added, as shown in Figure 9-52a. The link with a value of six is chosen. Now the set of candidate links consists of any link connected to this selected link (the two links with costs of 15 and 8, respectively), plus any connected to the starting point. All paths are examined, and the link added that gives the lowest total path length. In Figure 9-52b, two links are added. Note that the links added are not connected to the initially selected link. This would have given a total cost of 14 (6 plus 8) or 21 (6 plus 15), while the selected links give a lower path cost of 12. Now, the candidate links are those connected to any of the selected links or to the start point. Since all links from the start point have been selected, only those connected to candidate links are examined. Of these, the lowest cost path is added. The link with a cost of 8 that is attached to the initially selected link is chosen (Figure 9-52c). The candidate set expands accordingly, and is evaluated again. Verify that the links shown in Figure 9-52d and Figure 9-52e should be the next, cumulative low cost paths selected. This method is used until the destination is reached, and the least-cost path identified (Figure 9-53).

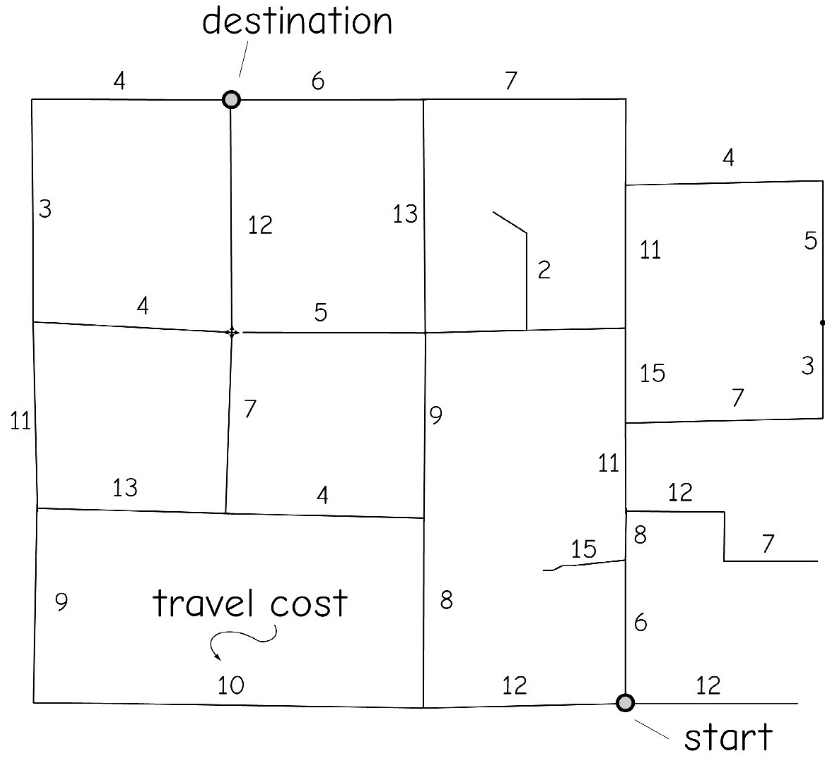


Figure 9-51: An example network. Start and destination centers and costs for link traversal are shown.

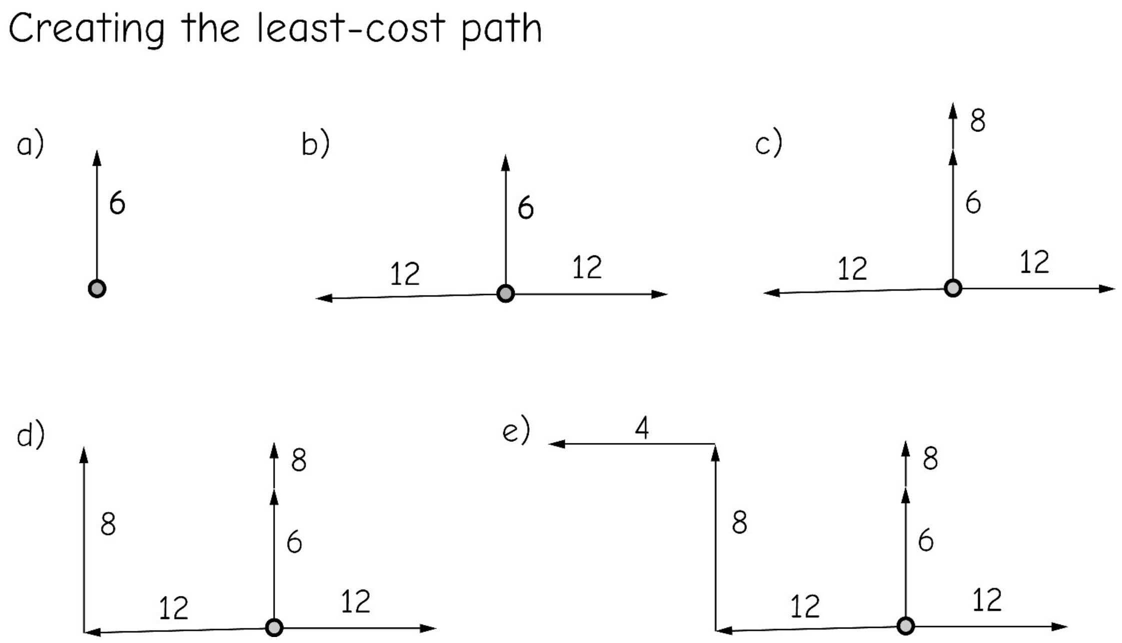


Figure 9-52: Steps in the identification of the least-cost path.

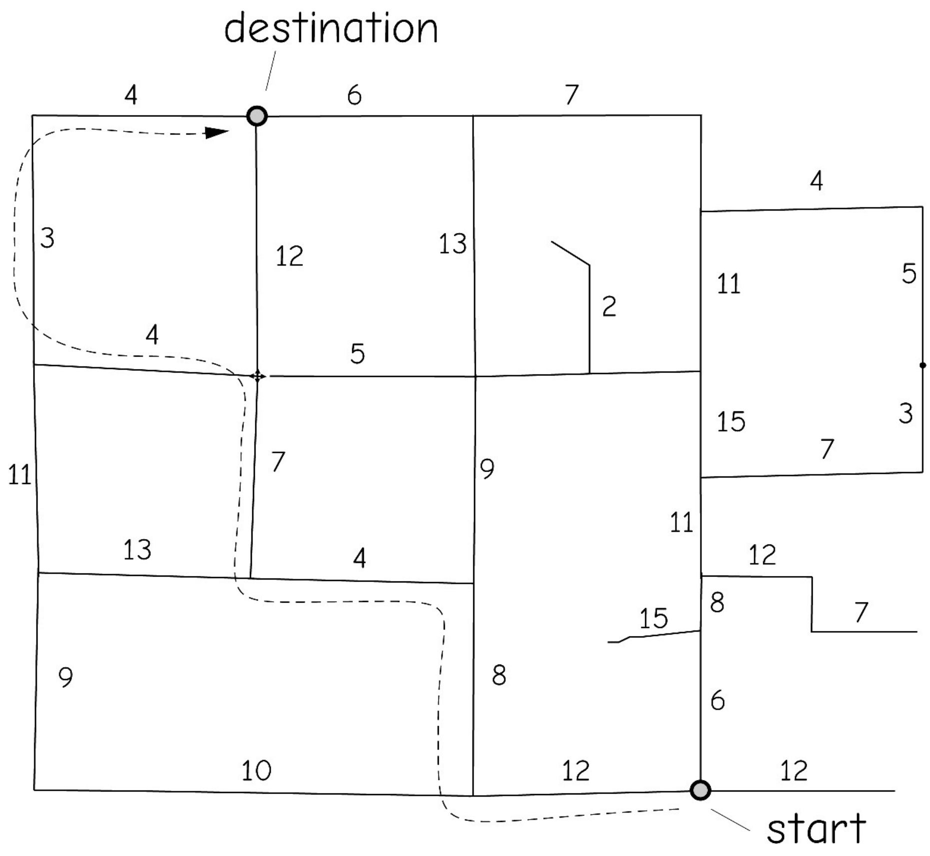


Figure 9-53: Least-cost path for the example route finding algorithm described in the text.

Many different pathfinding algorithms have been developed, most of which are much more sophisticated than the one described above. Note that the described pathfinding algorithm has a rapidly expanding number of links to evaluate at each step. Computational burdens increase accordingly. A subset of all possible candidate paths may be examined because it becomes too computationally time-consuming to examine all possible paths. Most pathfinding algorithms periodically review the total accumulated cost for each candidate path, and stop following some highest cost or least promising paths.

There are many variations on this route finding problem. There may be multiple centers that must be visited in a specific order, and carriers defined to transport specific amounts to or from centers. Centers may add to or subtract from a carrier; for example, some centers might represent houses with children, other centers may represent schools, and carriers represent buses that transport children. Houses must be visited to pick up children, but a bus has a fixed capacity. These children must be transported to the school, and there may be time constraints; for example, children cannot be picked up before 7 a.m. and must be at school by 7:55 a.m. Network-based route selection has been successfully used to solve these and related problems.

Resource allocation problems involve the apportionment of a network to centers. One or more allocation centers are defined in a network. Territories are defined for each of these centers. Territories encompass links or non-allocation centers in the network. These links or non-allocation centers are assigned to only one allocation center. The features are usually assigned to the nearest center, where distance is measured in time, length, or monetary units.

Resource allocation algorithms may be similar to route finding algorithms in that the distance out from each center is calculated along each path. Each center or arc is assigned to the nearest or least-cost center. The route finding method is exhaustive in resource allocation, in that all routes are pursued, not just the least-cost route. The routes are measured outward from each allocation center (Figure 9-54).

Variations on resource allocation include setting a center capacity. The center capacity sets an upper limit on resources that may be encompassed by a territory. Links are assigned to the nearest center, but once the capacity is reached, no more are added. Maximum distance also serves to limit the range of the territory from the center. Both of these restrictions may result in some unassigned areas, that is, portions of the network that are not allocated to a center.

Resource allocation analyses are used in many disciplines. School districts may use resource allocation to assign neighborhoods to schools. The type and number of dwellings in a district may be included as nodes on a network. The number of children along each link is added until the school capacity is reached. Resource allocation may also be used to define sales territories, or to determine if a new business should be located between existing businesses. If enough customers fall between the territories of existing business centers, a new business between existing business centers may be justified.

Traffic modeling is another oft-applied network analysis. Streets are represented by a network of interconnected arcs and nodes. Attributes associated with arcs define travel speed and direction. Attributes associated with nodes identify turns and the time or cost required for each turn. Illegal or impossible turns may be modeled by specifying an infinite cost. Traffic is placed in the network, and movement modeled. Bottlenecks, transit times, and underused routes may be identified, and this information used to improve traffic management or build additional roads.

Traffic modeling through networks is a subdiscipline in its own right. Due to the cost and importance of transportation and traffic management, a great deal of emphasis has been placed on efficient traffic management. Transportation engineers, computer

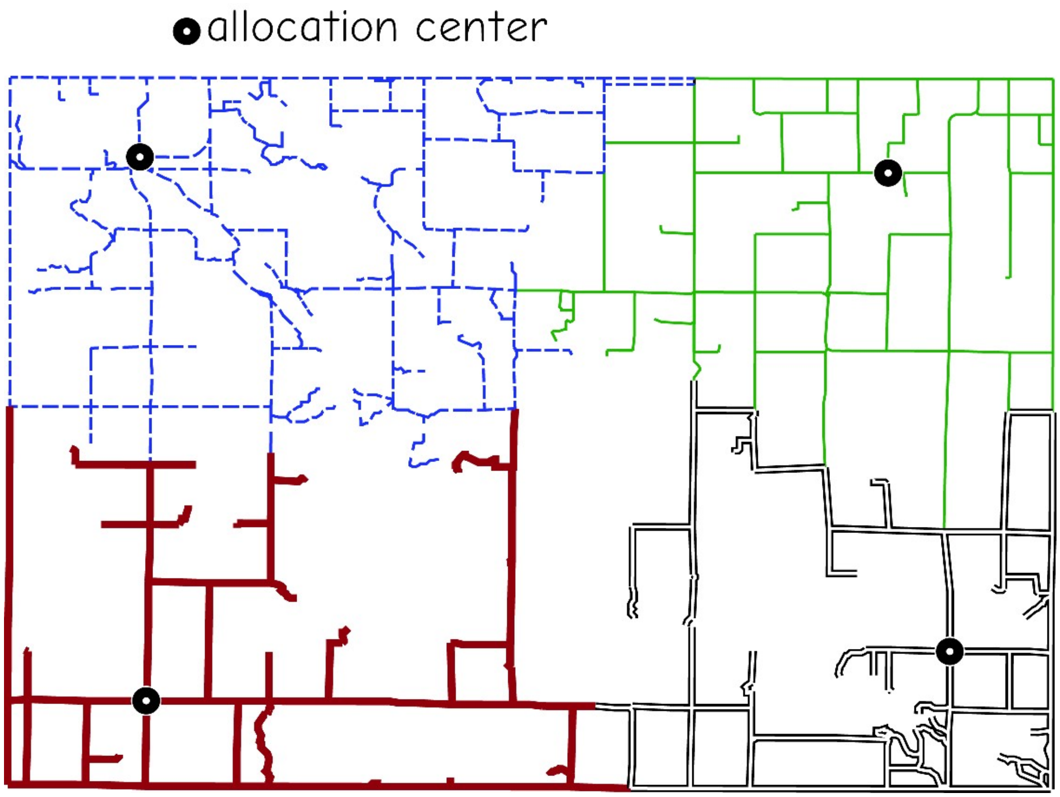


Figure 9-54: Allocation of network links to distinct centers. Network links or resources are assigned to the "nearest" center, where distance may be determined by physical distance, or by cost, travel time, or some other factor.

scientists, and mathematicians have been modeling traffic via networks for many years. An in-depth discussion of network analyses for traffic management may be found in literature listed at the end of this chapter.

Geocoding

Geocoding, also known as *linear referencing*, is another common application of spatial data networks. Geocoding is the process of spatially referencing point features based on the address of the feature and knowledge of an address range for the linear network. Geocoding is commonly applied for business sales, in marketing, in vehicle dispatch and delivery operations, and for organizing censuses and other government information gathering and dissemination activities.

Geocoding requires a set of addresses associated with a set of linear features. Typically, at least the starting and ending addresses for links in a network are known. These starting and ending addresses define an address range, and the range is assumed to linearly span the connecting line. Points on the line may be "geographically coded" (hence the name geocoding), in that given an address, we may calculate approximately where the address should occur on the network link (Figure 9-55).

Geocoded addresses are typically assumed to vary linearly along the link. The starting and ending address are assumed to be at the ends of the link. The estimated location of the geocoded address is based on a linear interpolation, beginning at the starting address and adding a length proportional to the address divided by the address range (Figure 9-55). The estimated location may be placed within the block or line segment.

Because geocoding only estimates address locations, these locations may contain substantial error. These errors may be larger than the error associated with the linear features along which the geocoded addresses are placed. Figure 9-56 illustrates some sources of error. Geocoding typically involves a regular, linear interpolation of an address across an address range. Address ranges are usually assigned ordinally, while the geocode is an interval estimate. In Figure 9-56a, address 250 is not halfway between 200 and 300, and address 240 takes up an entire block. This ordinal/interval mismatch may be particularly bad in rural areas, where development over a long time period may

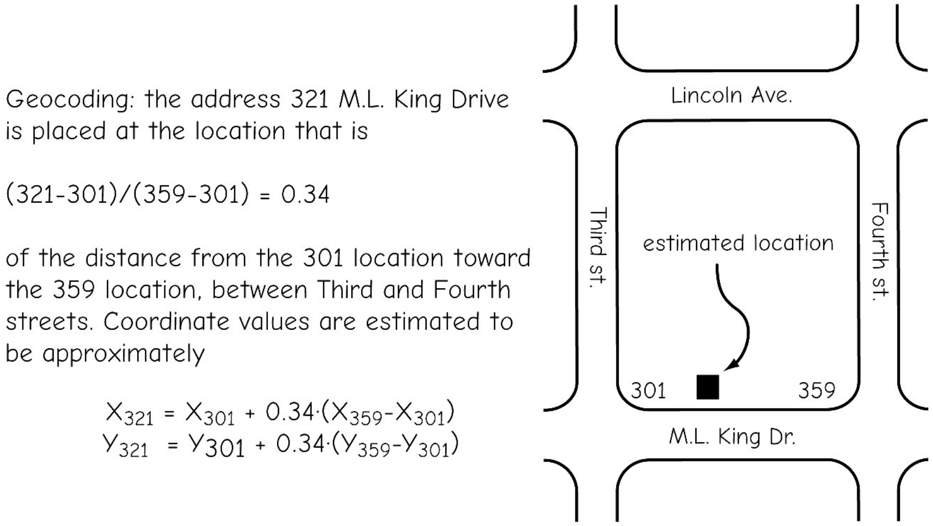


Figure 9-55: Geocoding is the process of estimating the location of an address based on knowledge of an address range along a linear feature. Here an address location is linearly interpolated along a city block, giving the approximate location of a building.

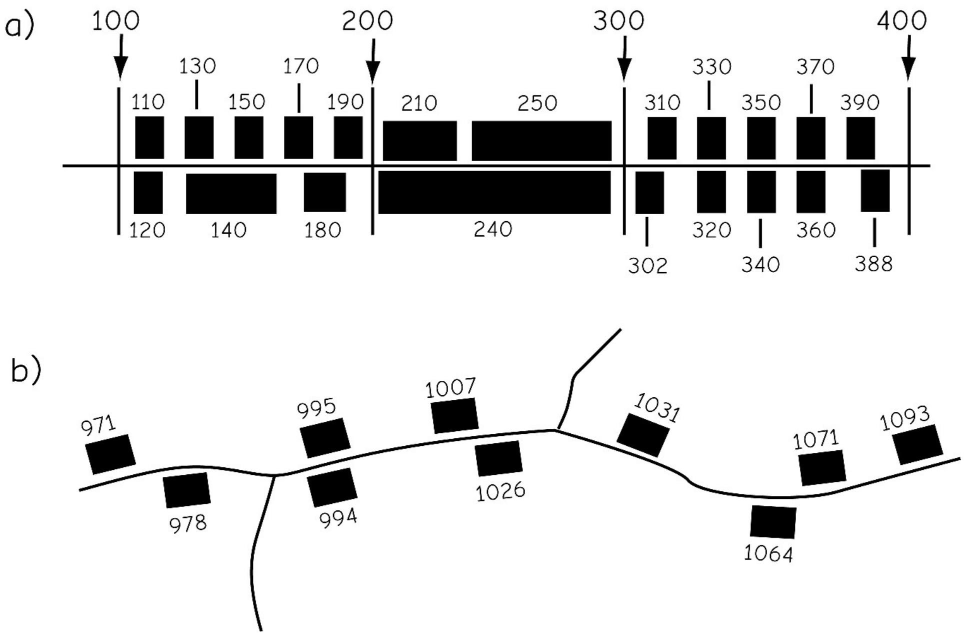


Figure 9-56: Idiosyncratic development may result in non-linear address sequences on the ground, so errors may result when geocoding is applied. Geocoding is typically applied with an assumption of a linear distribution of addresses across a range. When this is not true, as illustrated with address 250 in part a, above, or 1026 in part b, geocoded locations will be in error.

result in substantial nonlinear address arrangements. Figure 9-56b illustrates this, with address 1007 almost opposite address 1026, and numerous inconsistent intervals; for example, the 22 address units between 1071 and 1093 are separated by a shorter distance than the 12 address units between 995 and 1007. These nonlinear addresses can cause substantial confusion, so any application of geocoded data must allow for these inconsistencies, or the data must be evaluated and corrected.

Geocoding is often combined with network analyses to determine shortest path or time travels to a set of locations. Delivery locations may be generated from a list of orders to a business. The locations of these addresses are generated via geocoding. The locations may then be entered into a network search algorithm and the optimal route planned. Businesses save millions of dollars each year applying these basic spatial analyses.

Summary

Spatial analysis, along with map production, is one of the most important uses of GIS. Spatial analytical capabilities are often the reason we obtain GIS and invest substantial time and money to develop a working system. Any analytical operation we perform on our spatial or associated attribute data may be considered as spatial analysis.

Spatial operations are applied to input data and generate output data. Inputs may be one to many layers of spatial data, as well as nonspatial data. Outputs may also number from one to many. Operations also have a spatial scope, the area of the input data that contributes to output values. Scopes are commonly local, neighborhood, or global.

Selection and classification are among the most oft-used spatial data operations. A selection identifies a subset of the features in a spatial database. The selection may be based on attribute data, spatial data, or some combination of the two. Selection may apply set or Boolean algebra, and may combine these with analyses of adjacency, connectivity, or containment. A selected set may be classified in that variables may be changed or new variables added that reflect membership in the selected set.

Classifications may be assigned automatically, but the user should be careful in choosing the assignment. Equal-area, equal-interval, and natural breaks classifications are often used. The resulting classifications may depend substantially on the frequency histogram of the input data layer, particularly when outliers are present.

A dissolve operation is often used in spatial analysis. Dissolves are routinely applied after a classification, as they remove redundant boundaries that may slow processing.

Proximity functions and buffers are also commonly applied spatial data operations. These functions answer questions regarding distance and separation among features in the same or different data layers. Buffering may be applied to raster or vector data, and may be simple (with a uniform buffer distance), or complex (with multiple nested buffers or variable buffer distances).

Overlay involves the vertical combination of data from two or more layers. Both geometry (coordinates) and attributes are combined. Any combination of points, lines, and area features is possible, although overlays involving at least one layer of area features are most common. The results of an overlay usually take the lowest geometric dimension of the input layers.

Overlay sometimes creates gaps and slivers. These occur most often when a common feature occurs in two or more layers. These gaps and slivers may be removed by several techniques.

Network models may be temporally dynamic or static, but they are constrained to model the flow of resources through a connected set of linear and point features. Traffic flow, oil and gas delivery, or electrical networks are examples of features analyzed and managed with network models. Route finding, allocation, and flow are commonly modeled in networks.

Geocoding, or linear referencing, is used to calculate approximate locations along a linear segment when the endpoint addresses are known. Often used in census and delivery applications, geocoding works best when addresses are uniformly spaced across the segment. Because it is an approximation, geocoded locations are expected to sometimes be in error, and these errors are often more frequent in rural or sparsely addressed segments. Linear referencing may also be used to locate changes in linear feature characteristics, for example, road surface or accident locations.

Suggested Reading

Ahuja, R.K., Magnanti, T.L., Orlin, J.B. (1993). *Network Flows: Theory, Algorithms, and Applications.* Englewood Cliffs: Prentice Hall.

Aronoff, S. (1989). *Geographic Information Systems, A Management Perspective*. Ottawa: WDL Publications.

Batty, M., Xie, Y. (1994). Model structures, exploratory spatial data analysis, and aggregation. *International Journal of Geographical Information Systems*, 8:291–307.

Bonham-Carter, G.F. (1996). *Geographic Information Systems for Geoscientists: Modelling with GIS*. Ottawa: Pergamon.

Carver, S.J. (1991). Integrating multi-criteria evaluation with geographical information systems. *International Journal of Geographical Information Systems*, 5:321–340.

Chou, Y. H. (1997). *Exploring Spatial Analysis in Geographic Information Systems*. Albuquerque: Onword Press.

Cliff, A.D., Ord, J.K. (1981). *Spatial Processes: Models and Applications*. London: Pion.

Cooper, L. (1963). Location-allocation problems. *Operations Research*, 11:331–342.

Dale, P. (2005). *Introduction to Mathematical Techniques Used in GIS*. Boca Raton: CRC Press.

Daskin, M.S. (1995). *Network and Discrete Location — Models, Algorithms, and Applications.* New York: Wiley.

DeMers, M. (2000). *Fundamentals of Geographic Information Systems* (2nd ed.). New York: Wiley.

Heuvelink, G.B.M., Burrough, P.A. (1993). Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems*, 7:231–246.

Laurini, R., Thompson, D. (1992). *Fundamentals of Spatial Information Systems*. London: Academic Press.

Lombardi, J., Stern, E., Clarke, G. (2015). *Applied Spatial Modelling and Planning*. London: Routledge Press.

Malczewski, J. (1999). *GIS and Multicriteria Decision Analysis*. New York: Wiley.

Martin, D. (1996). *Geographical Information Systems and their Socio-economic Applications* (2nd ed.). London: Routledge.

McMaster, S., McMaster, R.B. (2002). Biophysical and human-social applications. J.D. Bossler (Ed.), *Manual of Geospatial Science and Technology*. London: Taylor and Francis.

Monmonier, M. (1993). *How To Lie With Maps*. Chicago: University of Chicago Press.

National Research Council of the National Academies (2006). *Beyond Mapping: Meeting National Needs Through Enhanced Geographic Information Science.* Washington D.C.: The National Academies Press.

Openshaw, S., Taylor, P. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. N. Wrigley (Ed.) *Statistical Applications in the Spatial Sciences*. London: Pion.

Smith, M.J.de, Goodchild, M.F., Longley, P.A. (2007). *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*. Leicester: Winchelsea Press.

Steinitz, C., Jordan, L. (1976). Hand-drawn overlays: Their history and perspective uses. *Landscape Architecture*, 56:146–157.

Stillwell, J.A., Clarke, G. (2004). *Applied GIS and Spatial Analysis*. New York: Wiley.

Worboys, M.F., Duckham, M. (2004). *GIS: A Computing Perspective* (2nd ed.). Boca Raton: CRC Press.

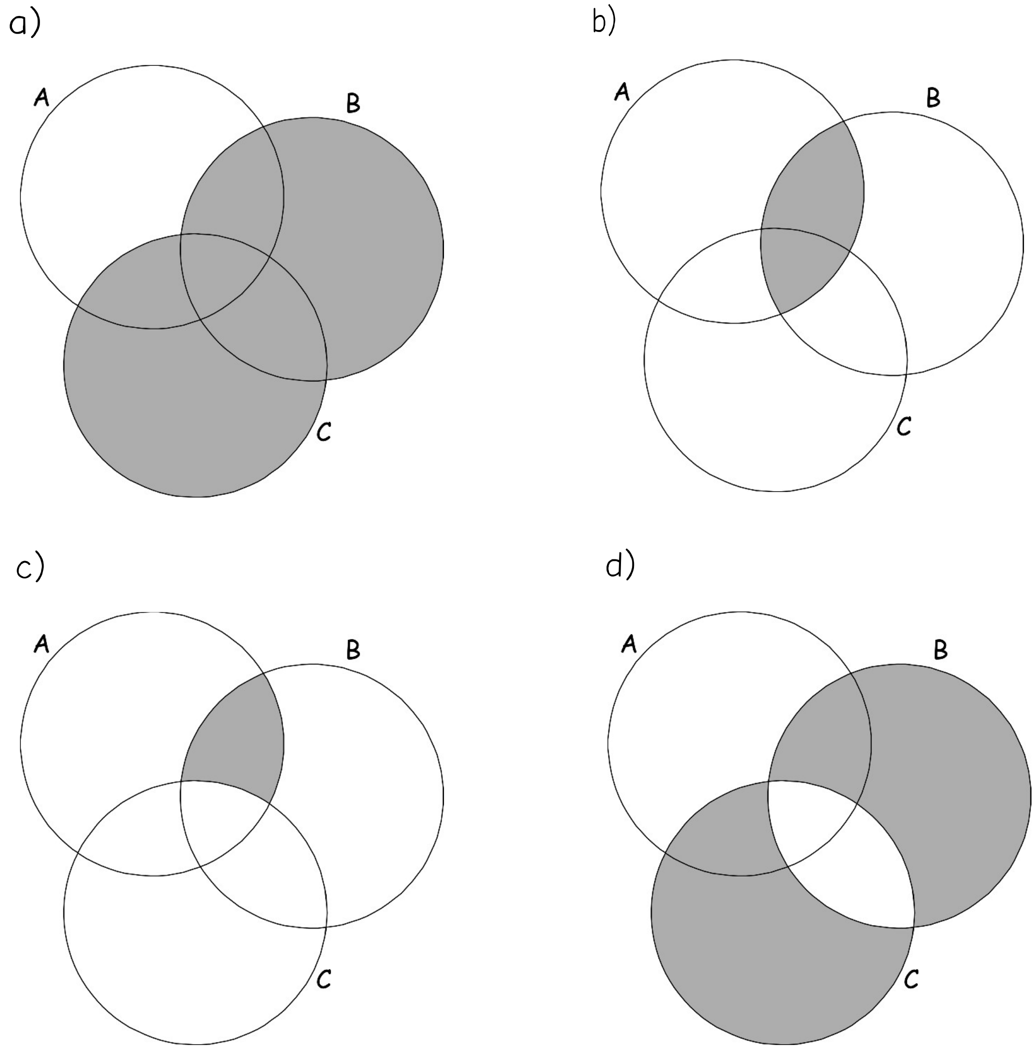
Study Problems and Questions

9.1 - Define and give examples of local, neighborhood, and global spatial operations.

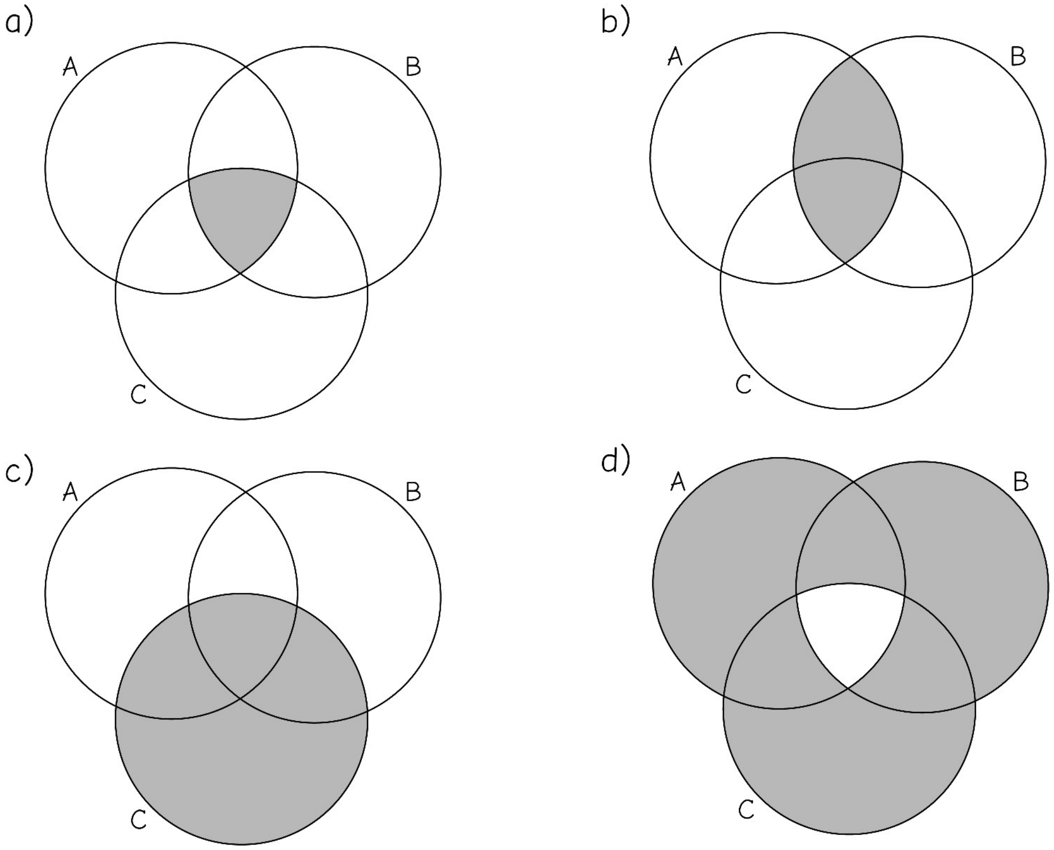
9.2 - Describe selection operations.

9.3 - Describe set and Boolean algebra.

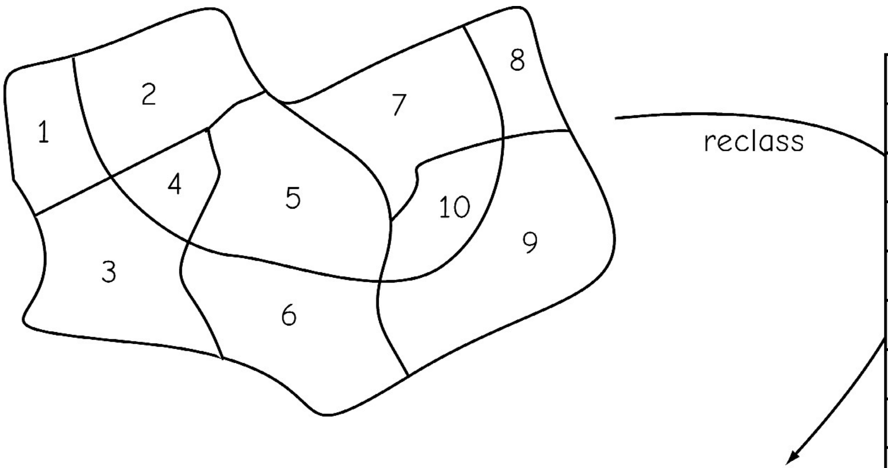
9.4 - Write the simplest Boolean expressions that result in the grey area selections:



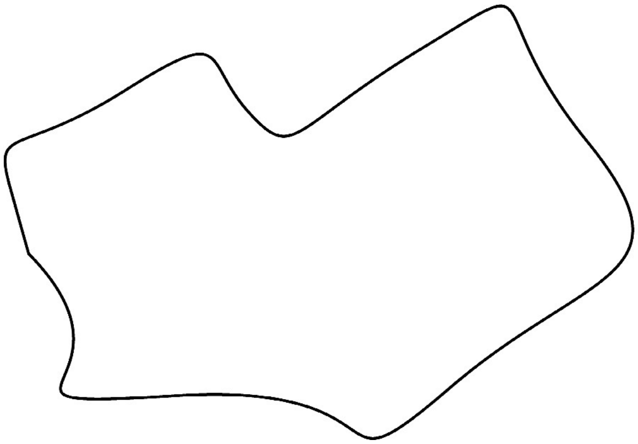
9.5 - Write the simplest Boolean expressions that result in the grey area selection:



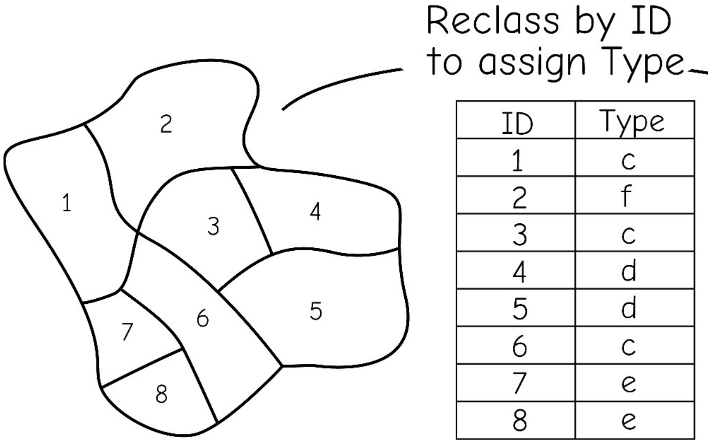
9.6 - Perform the following reclassification:



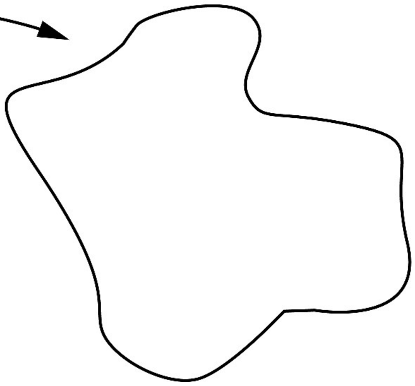
ID	catg.
1	A
2	A
3	B
4	A
5	B
6	B
7	A
8	A
9	C
10	C



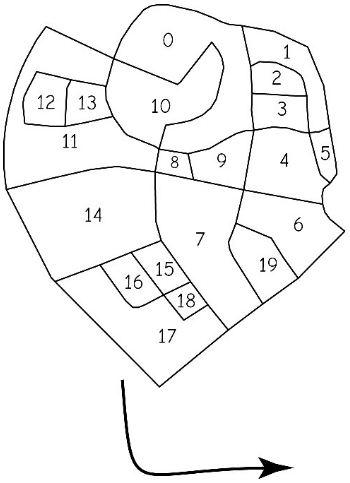
9.7 - Perform the following reclassification:



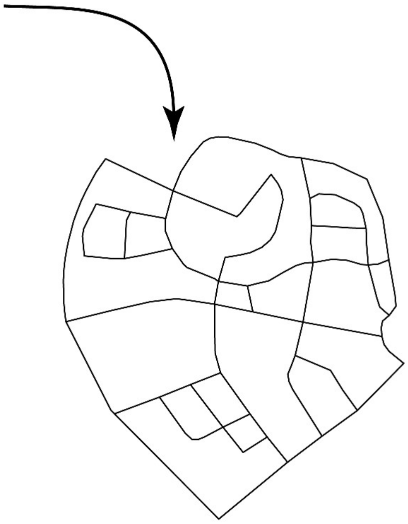
ID	Type
1	c
2	f
3	c
4	d
5	d
6	c
7	e
8	e



9.8 - Reclassify the following polygons, according to the column Area, into small (<18,000), medium (18,000 to 45,000), and large (> 45,000).



ID	Area	Type
0	100,074	low
1	30,240	high
2	14,228	high
3	17,770	high
4	43,923	low
5	8,688	high
6	46,169	mid
7	87,915	low
8	7,762	high
9	23,318	mid
10	62,201	low
11	104,749	mid
12	19,036	mid
13	15,239	high
14	111,718	low
15	14,854	high
16	16,948	high
17	77,985	low
18	8,281	high
19	27,437	mid

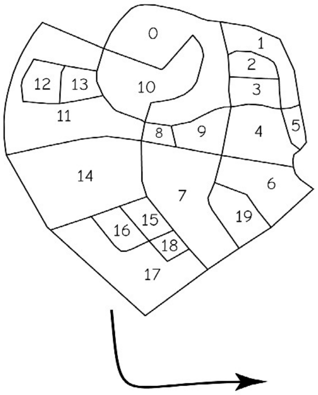


9.9 - List and describe three different classification methods.

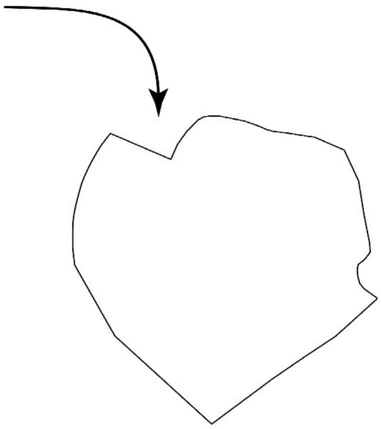
9.10 - What is the modifiable area unit problem (MAUP)? Why is it important? What is the zone effect, and what is the area effect?

9.11 - What is a dissolve operation? What are they typically used for?

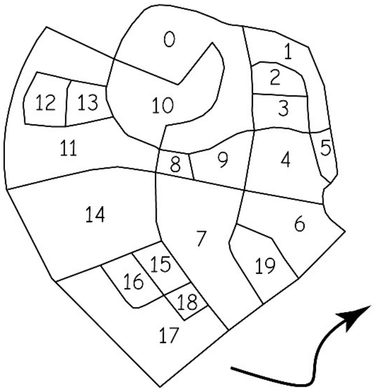
9.12 - Perform a dissolve operation on the variable Type for the layer depicted below:



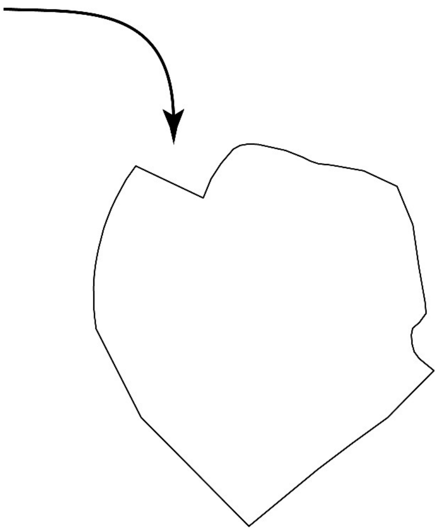
ID	Area	Type
0	100,074	low
1	30,240	high
2	14,228	high
3	17,770	high
4	43,923	low
5	8,688	high
6	46,169	mid
7	87,915	low
8	7,762	high
9	23,318	mid
10	62,201	low
11	104,749	mid
12	19,036	mid
13	15,239	high
14	111,718	low
15	14,854	high
16	16,948	high
17	77,985	low
18	8,281	high
19	27,437	mid



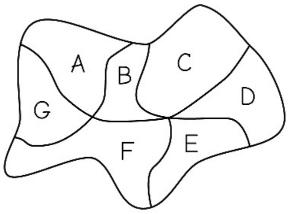
9.13 - Perform a dissolve operation on the variable Type for the layer depicted below:



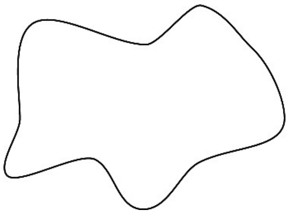
ID	Area	Type
0	100,074	high
1	30,240	low
2	14,228	high
3	17,770	high
4	43,923	low
5	8,688	high
6	46,169	low
7	87,915	low
8	7,762	high
9	23,318	mid
10	62,201	low
11	104,749	mid
12	19,036	mid
13	15,239	high
14	111,718	low
15	14,854	high
16	16,948	high
17	77,985	high
18	8,281	high
19	27,437	mid



9.14 - Draw the resultant polygon boundaries and complete the table in a dissolve operation that calculates the sum of Count, based on Class. Label each polygon starting lowercase a, b, c..., and enter the label in NewID in the output table for the corresponding row.

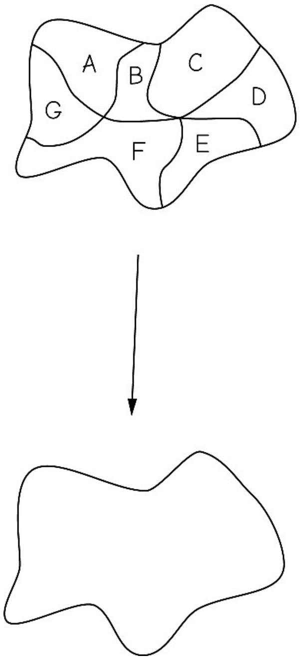


ID	Class	Count	Type	Cost
A	1	11	Farm	1,000
B	2	9	Farm	900
C	1	3	Ranch	1,100
D	1	4	Suburb	200
E	2	21	Suburb	700
F	2	14	Farm	800
G	1	6	Ranch	1,200



NewID	Class	Count

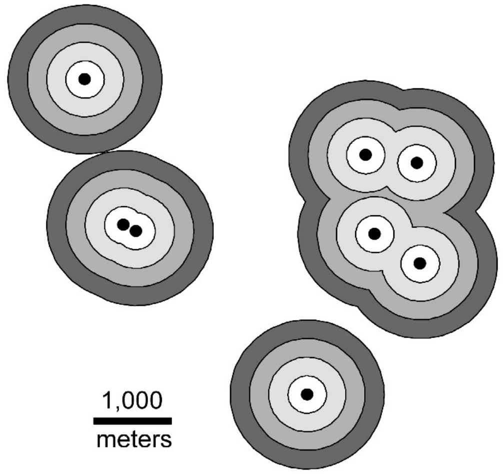
9.15 - Draw the resultant polygon boundaries and complete the table in the dissolve operation that calculates the average of `Cost`, based on `Type`. Label each output polygon starting lowercase a, b, c..., and enter the label in `NewID` in the output table for the corresponding row.



ID	Class	Count	Type	Cost
A	1	11	Farm	1,000
B	2	9	Farm	900
C	1	3	Ranch	1,100
D	1	4	Suburb	200
E	2	21	Suburb	700
F	2	14	Farm	800
G	1	6	Ranch	1,200

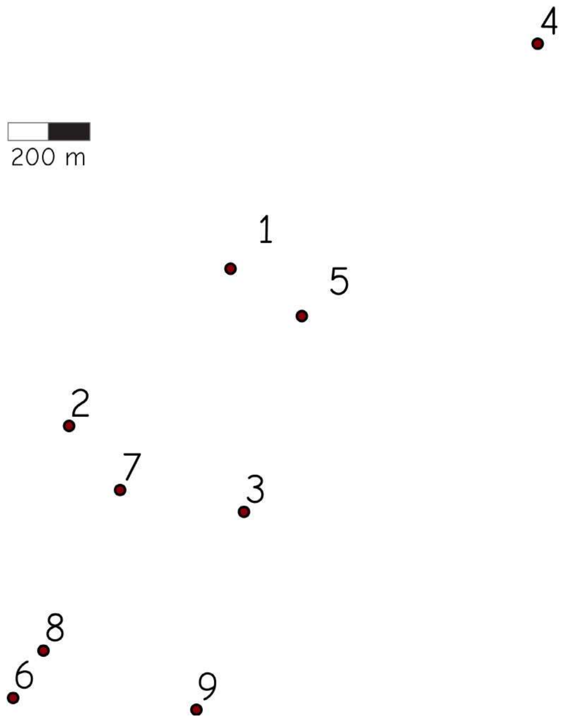
NewID	Class	Count

9.16 - Select the most appropriate characteristics for the buffer below. Is it simple, multi-distance, or variable distance? Does it retain or dissolve intersections? Is it interior or exterior?

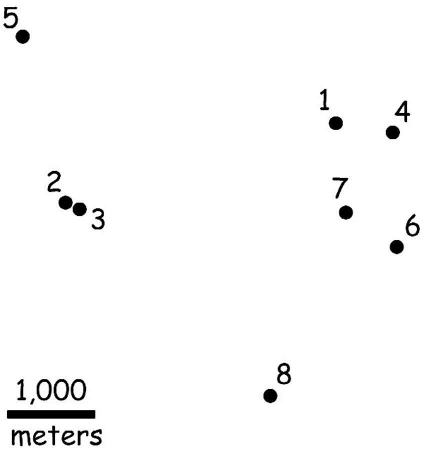


9.17 - Sketch out the output from a variable distance buffer applied to the set of points shown below. Draw output buffers that dissolve the boundaries between areas that fall within multiple buffers.

ID	dist.
1	100
2	200
3	150
4	300
5	100
6	800
7	200
8	100
9	150



9.18 - Sketch out the output from a variable distance buffer applied to the set of points shown below. Dissolve boundaries for intersecting buffers.

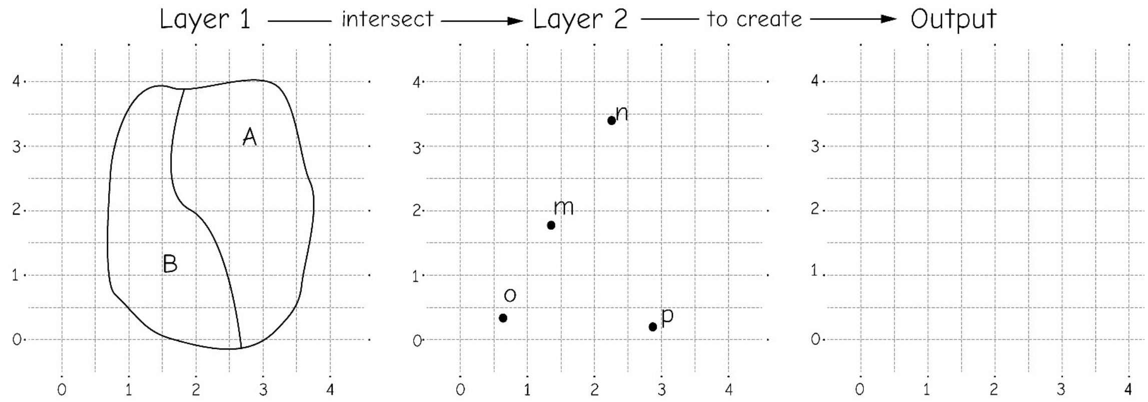


ID	distance
1	250
2	0
3	1500
4	500
5	0
6	500
7	250
8	250

9.19 - How are raster proximity functions different from vector proximity functions?

9.20 - Why are output features in vector overlay typically set to the minimum dimensional order (point, line, or polygon) of the input features?

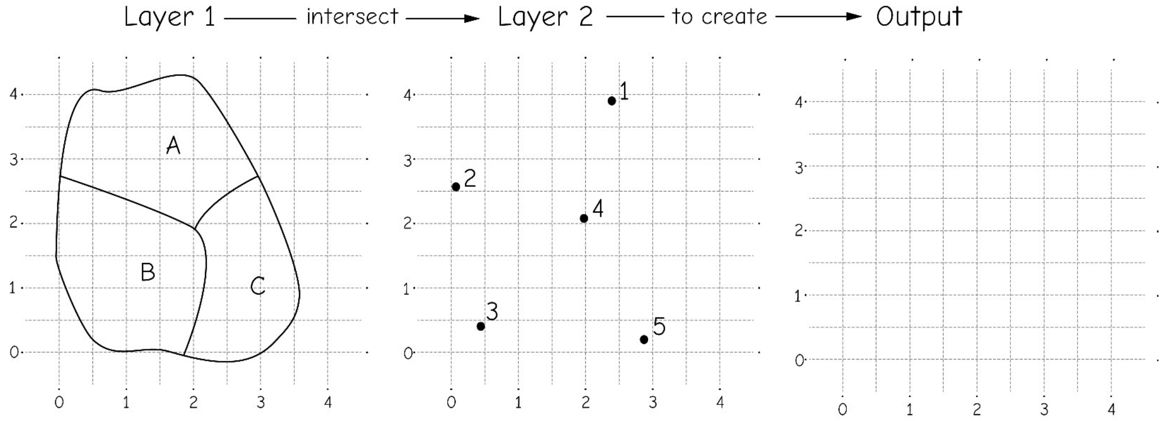
9.21 - Complete the table for the vector point overlay shown below:



ID	Lndcov	county
A	forest	Wallowa
B	urban	Lincoln

ID	name	type
m	Boulder	CZy
n	Butte	YaBo
o	Rockville	Sqr
p	Ophir	LeKr

9.22 - Complete the table for the vector point overlay shown below:

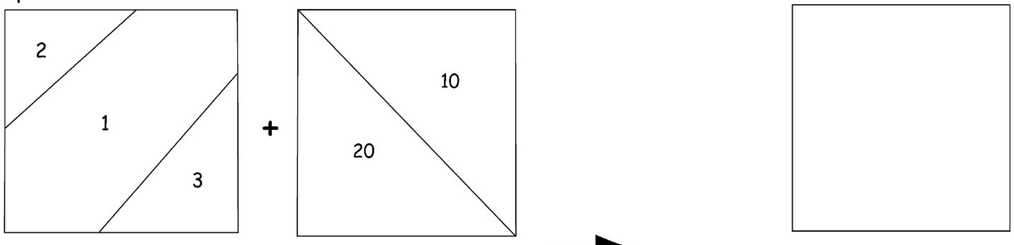


ID	code	stts
A	Grs	fee
B	Wtl	lease
C	Urb	TxDf

ID	ctgry
1	FtTyr
2	XC
3	STrk
4	Exp
5	CFbr

9.23 - Sketch both the output polygons and the resultant attribute table from the overlay shown below:

Spatial Data

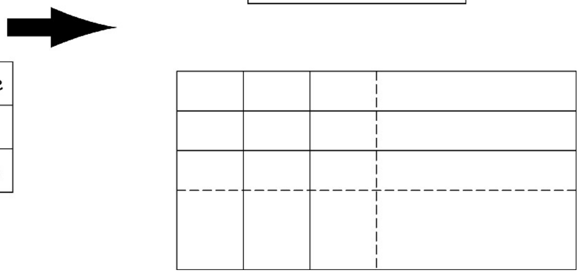


Tables

ID	type	year
1	fee	1902
2	lease	2007
3	common	1957

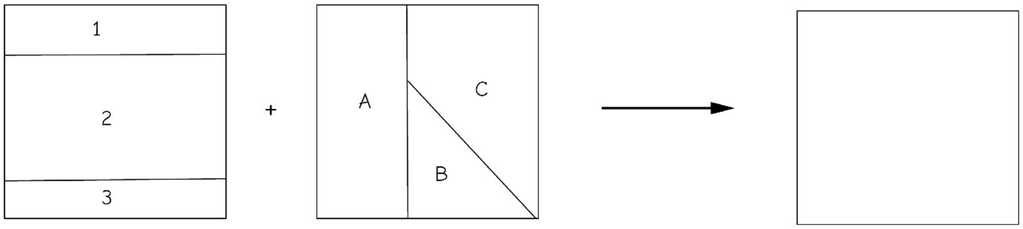
\+

ID	county	status	value
10	Polk	public	2.9
20	Tule	private	1.2M



9.24 - Sketch both the output polygons and the resultant attribute table from the overlay shown below:

Spatial Data

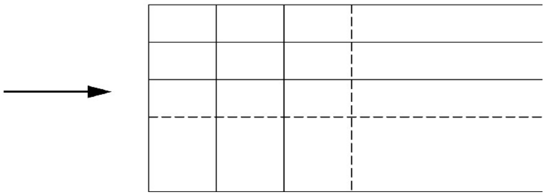


Tables

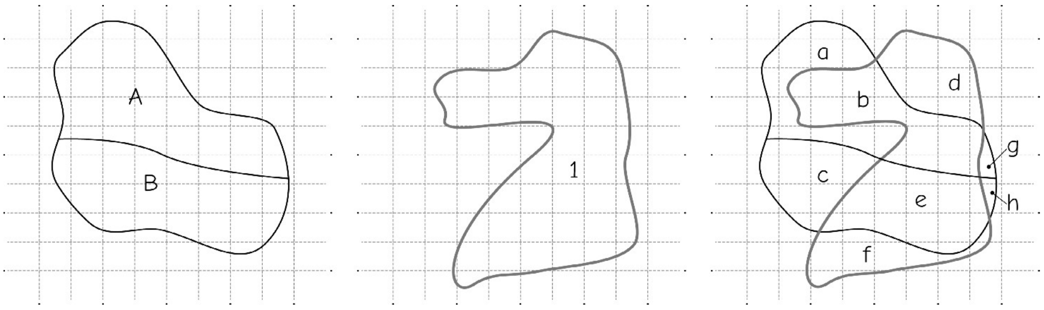
ID	no	own
1	1215	E. Nuff
2	1863	U. Kahn
3	1919	I. Cee

\+

ID	Ward	Rep	pop.
A	11	Hanks	496
B	2	Ryan	187
C	9	Gump	2,312



9.25 - Complete the table in the polygon union diagrammed below:

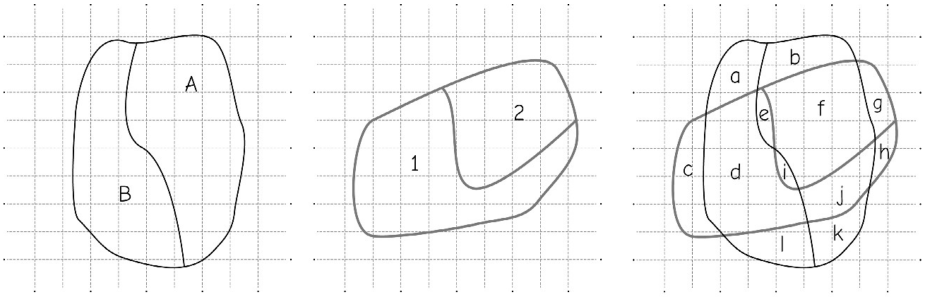


ID	status	zoning
A	private	high dense
B	public	park

ID	domveg	year	rcode
A	forest	1977	247A

ID				
a				
b				
c				
d				
e				
f				
g				
h				

9.26 - Complete the table in the polygon union diagrammed below:



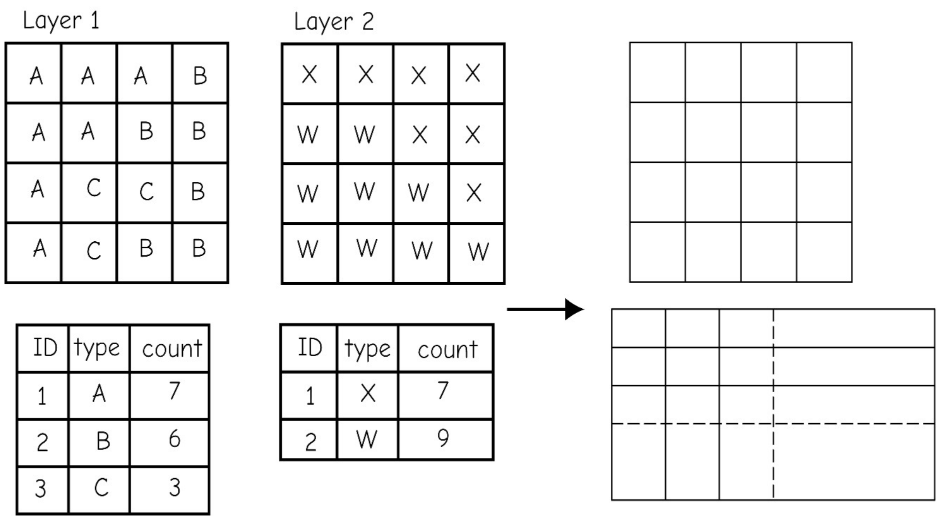
ID	Lndcov	county
A	forest	Wallowa
B	urban	Lincoln

ID	type	txtr
1	Evern	sand
2	Gillard	clay

ID				
a				
b				
c				
d				
e				
f				
g				
h				
i				
j				
k				
l				

9.27 - What is the sliver problem in vector layer overlay? How might this problem be resolved?

9.28 - Sketch the output of the raster overlay shown below, providing both cell values and the output table with ID and count variables.



Layer 1

A	A	A	B
A	A	B	B
A	C	C	B
A	C	B	B

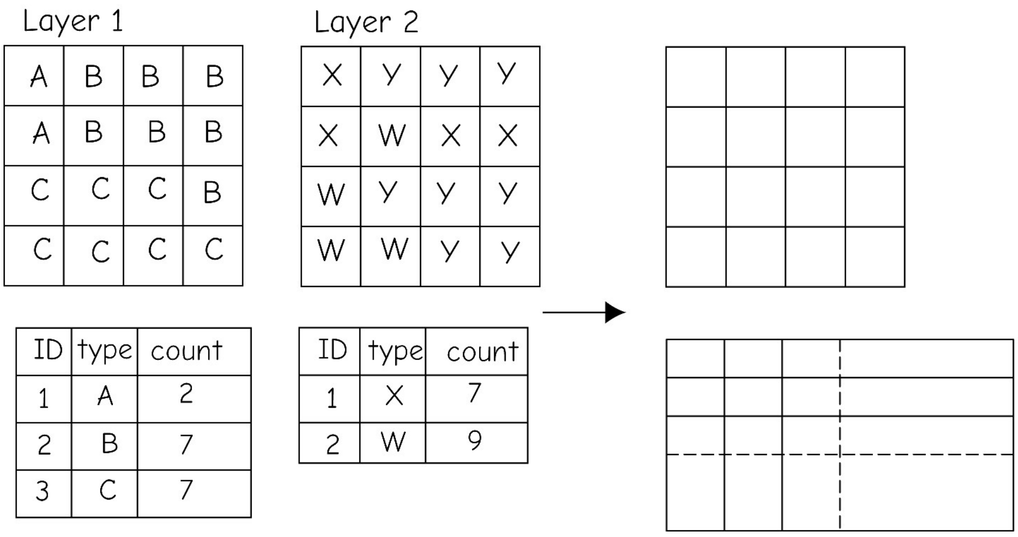
Layer 2

X	X	X	X
W	W	X	X
W	W	W	X
W	W	W	W

ID	type	count
1	A	7
2	B	6
3	C	3

ID	type	count
1	X	7
2	W	9

9.29 - Sketch the output of the raster overlay shown below, providing both cell values and the output table with ID and count variables.



Layer 1

A	B	B	B
A	B	B	B
C	C	C	B
C	C	C	C

Layer 2

X	Y	Y	Y
X	W	X	X
W	Y	Y	Y
W	W	Y	Y

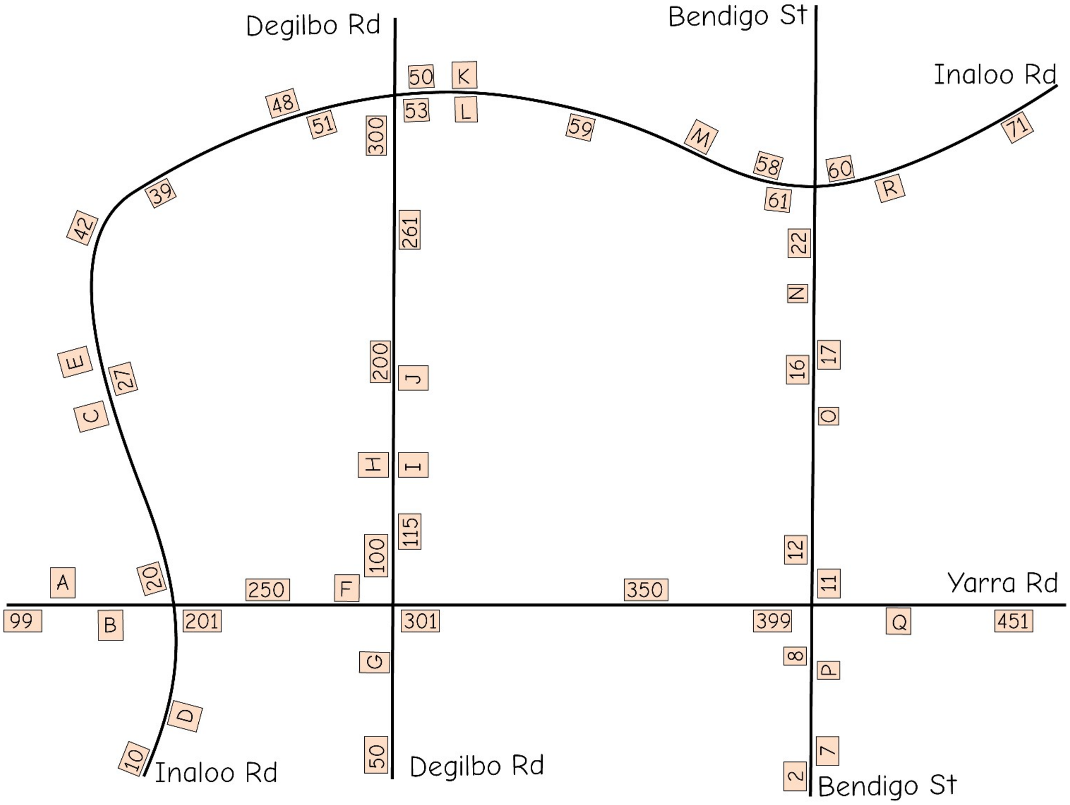
ID	type	count
1	A	2
2	B	7
3	C	7

ID	type	count
1	X	7
2	W	9

9.30 - Describe/define network models. What distinguishes them from other spatial or temporal models?

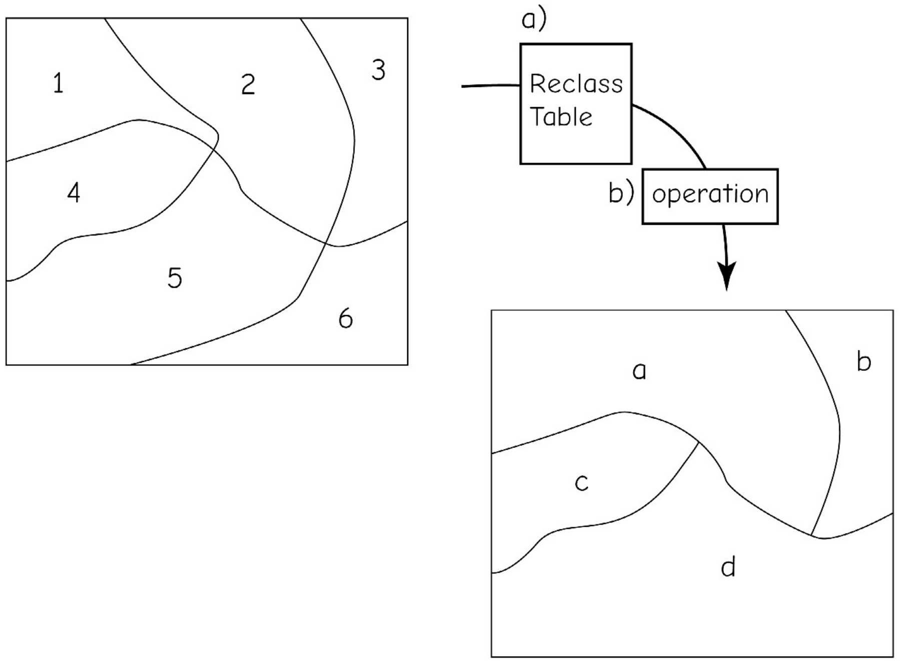
9.31 - What are the common uses for network models? Why are these models so important?

9.32 - Calculate and record the geocoded address for the boxes labeled A, C, E, G, I, K, M, O, and Q in the figure below:

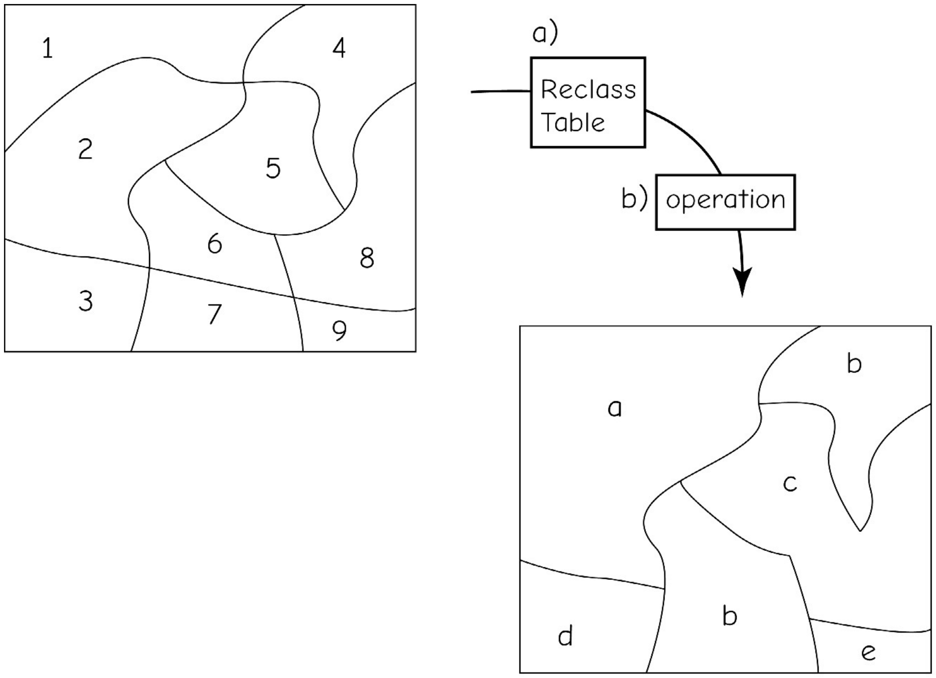


9.33 - Use the figure above to calculate and record the geocoded address for the boxes labeled B, D, F, H, J, L, N, P, and R.

9.34 - If you start with the layer on the left, and wish to create the layer on the right, what would be the table you would use for the reclass operation at a), and what single spatial operation would you use in b) to obtain the desired result?



9.35 - If you start with the layer on the left, and wish to create the layer on the right, what would be the table you would use for the reclass operation at a), and what single spatial operation would you use in b) to obtain the desired result?



10 Topics in Raster Analysis

Introduction

Raster analyses range from the simple to the complex, largely due to the early invention, simplicity, and flexibility of the raster data model. Rasters are based on two-dimensional arrays, data structures supported by many of the earliest programming languages. The raster row and column format is among the easiest to specify in computer code, thereby encouraging modification, experimentation, and the creation of new raster operations. Raster cells can store nominal, ordinal, or interval/ratio data, so a wide range of variables may be represented. Complex constructs may be built from raster data, including networks of connected cells, or groups of cells to form areas.

The flexibility of raster analyses has been amply demonstrated by the wide range of problems they help solve. Raster analyses may predict the fate of pollutants in the atmosphere, disease spread, animal migration, and crop yields. Time varying and wide area phenomena are often analyzed using raster data, particularly when remotely sensed inputs are available. Raster analyses are applied to a range of scales, from fine grained problems, for example, by the U.S. Environmental Protection Agency in hazard analysis of Superfund sites, to global-scale estimates of forest growth. Local, state, and regional organizations have used raster analyses at many scales in between.

Numerous research projects have embellished the basic raster data structure, and developed a general set of raster tools for spatial data analyses. Universities have developed raster analysis packages for research over the past six decades. Commercial raster GIS software has been created by a number of companies.

The long history of raster analyses has resulted in a set of tools that should be understood by every GIS user. Many of these tools have a common conceptual basis and they may be adapted to several types of problems. In addition, specialized raster analysis methods have been developed for less frequently encountered problems. The GIS user may more effectively apply raster data analysis if she understands the underlying concepts and has become acquainted with a range of raster analysis methods.

Map Algebra

Map algebra is the cell-by-cell combination of raster data layers. The combination entails applying a set of local and neighborhood functions, and to a lesser extent global functions, to raster data.

The concept of map algebra is based on the simple, flexible, and useful raster grid structure. Simple operations may be applied to each grid cell. Further, raster layers may be combined through operations such as layer addition, subtraction, and multiplication.

Map algebra entails operations applied to one or more raster data layers. *Unary* operations apply to one data layer. *Binary* operations apply to two data layers, and higher-order operations may involve many data layers.

A simple unary operation applies a function to each cell in an input raster layer, and records a calculated value to the corresponding cell in an output raster. Figure 10-1a illustrates the multiplication of a raster by a scalar (a single number). In this example the raster is multiplied by two. This might be denoted by the equation:

$$\text{Outlayer} = \text{Inlayer} * 2 \qquad (10.1)$$

Each cell value of In_layer is multiplied by 2, and the result placed in the corresponding cell in Outlayer. Other unary functions are applied in a similar manner; for example, each cell may be raised to an exponent, divided by a fixed number, or converted to an absolute value.

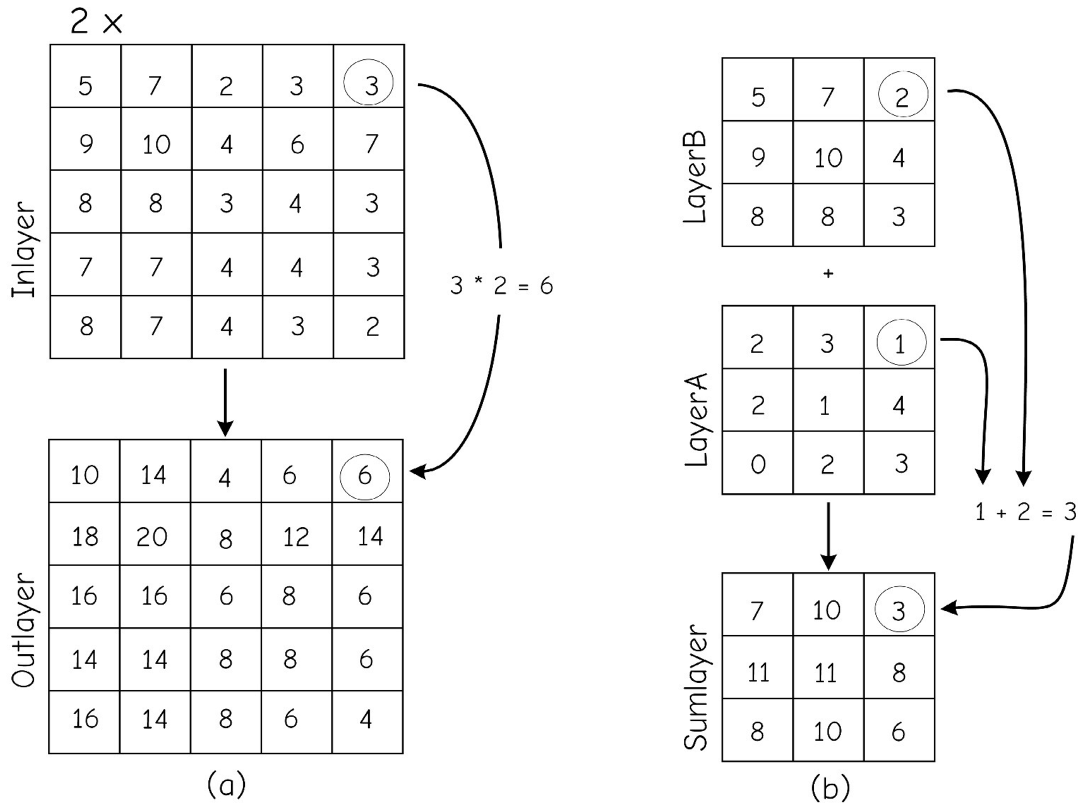


Figure 10-1: An example of raster operations. On the left side (a), each input cell is multiplied by the value 2, and the result stored in the corresponding output location. The right side (b) of the figure illustrates layer addition.

Binary operations also involve cell-by-cell application of operations or functions, but they combine data from two raster layers. Addition of two layers might be specified by:

Sumlayer = LayerA + LayerB (10.2)

Figure 10-1b illustrates this raster addition operation. Each value in LayerA is added to the value found in the corresponding cell in LayerB. These values are then placed in the appropriate raster cell of Sumlayer. The cell-by-cell addition is applied for the area covered by both LayerA and LayerB, and the results are placed in Sumlayer.

Note that in our example LayerA and LayerB have the same extent – they cover the same area. This may not always be true. When layer extents differ, most GIS software will either restrict the operation to the area where input layers overlap, or place a *null* or a “missing data” number to cells where input data are lacking. This number acts as a flag, indicating there are no results. It is often a number, such as –9999, that will not occur from an operation, but any placeholder may be used as long as the software and users understand the placeholder indicates no valid data are present.

Incompatible raster cell sizes cause ambiguities when raster layers are combined (Figure 10-2). This problem was described briefly in the previous chapter, and is illustrated with an additional example here. Consider cell A in Figure 10-2 when Layer1 and Layer2 are combined in a raster operation. Several cells in Layer2 correspond to cell A in Layer1. If these two layers are added, there are likely to be several different input values for Layer2 corresponding to one input value for Layer1. The problem is compounded in cell B, because a portion of the cell is not defined for Layer2. It falls outside the layer boundary. Which Layer2 value should be used in a raster operation? Is it best to choose only the values in Layer2 from the

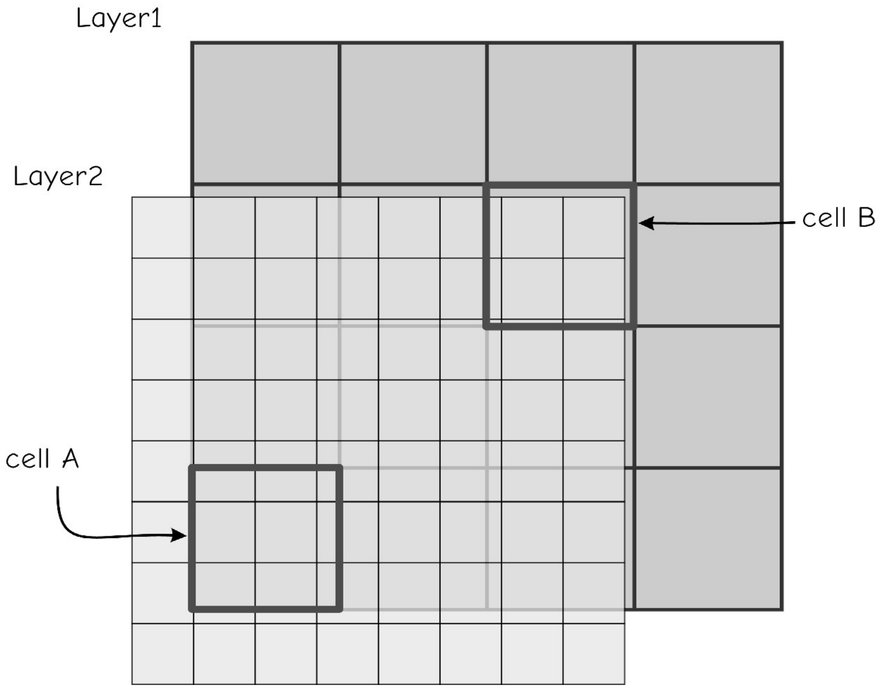


Figure 10-2: Incompatible cell sizes and boundaries confound multi-layer raster operations. This figure illustrates ambiguities in selecting input cell values from Layer2 in combination with Layer1. Multiple full and partial cells may contribute values for an operation applied to cell A. A portion of cell B is undefined in Layer2. These ambiguities are best resolved by resampling prior to layer combination, but if not done explicitly, the software may be written to do so automatically, sometimes with unintended results.

cells with complete overlap, or to use the median number, the average cell number, or some weighted average? This ambiguity will arise whenever raster data sets are not aligned or have incompatible cell sizes. While the GIS software may have a default method for choosing the "best" input when cells are different sizes or do not align, these decisions are best controlled by the human analyst prior to the application of the raster operation. The analyst may resample the data into a compatible coordinate system, using transformation and resampling methods described in Chapter 4.

As with vector operations, raster operations may be categorized as local, neighborhood, or global (Figure 10-3). Local operations use only the data in a single cell to calculate an output value. Neighborhood operations use data from a set of cells, and global operations use all data from a raster data layer.

The concepts of local and neighborhood operations are more uniformly specified with raster data than with vector data. Cells within a layer have a uniform size, so a local operation has a uniform input area. In contrast, vector areas represented by polygons may have vastly different areas. Irregular polygonal boundaries cover differing areas and have differing footprints; for example, the local area defined by Alaska is different than the local area defined by Rhode Island. A local operation in a given raster is uniform in that it specifies a particular cell size and dimension.

Neighborhood operations in raster data sets are also more uniformly defined than in vector data sets. A neighborhood may be defined by a fixed number of cells in a specific arrangement; for example, the neighborhood might be defined as a cell plus the eight surrounding cells. This neighborhood has a uniform area and dimension for most of the raster, with some minor adjustments needed near the edges of the raster data layer. Vector neighborhoods may depend not only on the shape and size of the target feature, but also on the shape and sizes of adjacent vector features.

Global operations in map algebra may produce uniform output, or they may produce different values for each raster cell. Global operations that return a uniform value are in effect returning a single number that is placed in every cell of the output layer. For example, the global maximum function for a layer might be specified as:

```
Out_num = globalmax(In_layer)        (10.3)
```

This would assign a single value to Out_num. The value would be the largest number found when searching all the cells of In_layer. This "collapsing" of data from a two-dimensional raster may reduce the map algebra to scalar algebra. Many other functions return a single global value placed in every cell for a layer, for example, the global mean, maximum, or minimum.

Global operations are at times quite useful. Consider an analysis of regional temperature. We may wish to identify the areas where daily maximum temperatures were warmer this year than the highest regional temperature ever recorded. This analysis might help us to identify the extent of a warming trend. We would first apply a maximum function to all previous yearly weather records. This would provide a scalar value, a single number representing the regional maximum temperature. We would then compare a raster data set of maximum temperature for each day in the current year against the "highest ever" scalar. If the value for a day were higher than the regional maximum, we would output a flag to a cell location. If it were not, we would output a different value. The final output raster would provide a map of the cells exceeding the previous regional maximum. Here we use a global operation first to create our single scalar value (highest regional temperature). This scalar is then used in subsequent operations that output raster data layers.

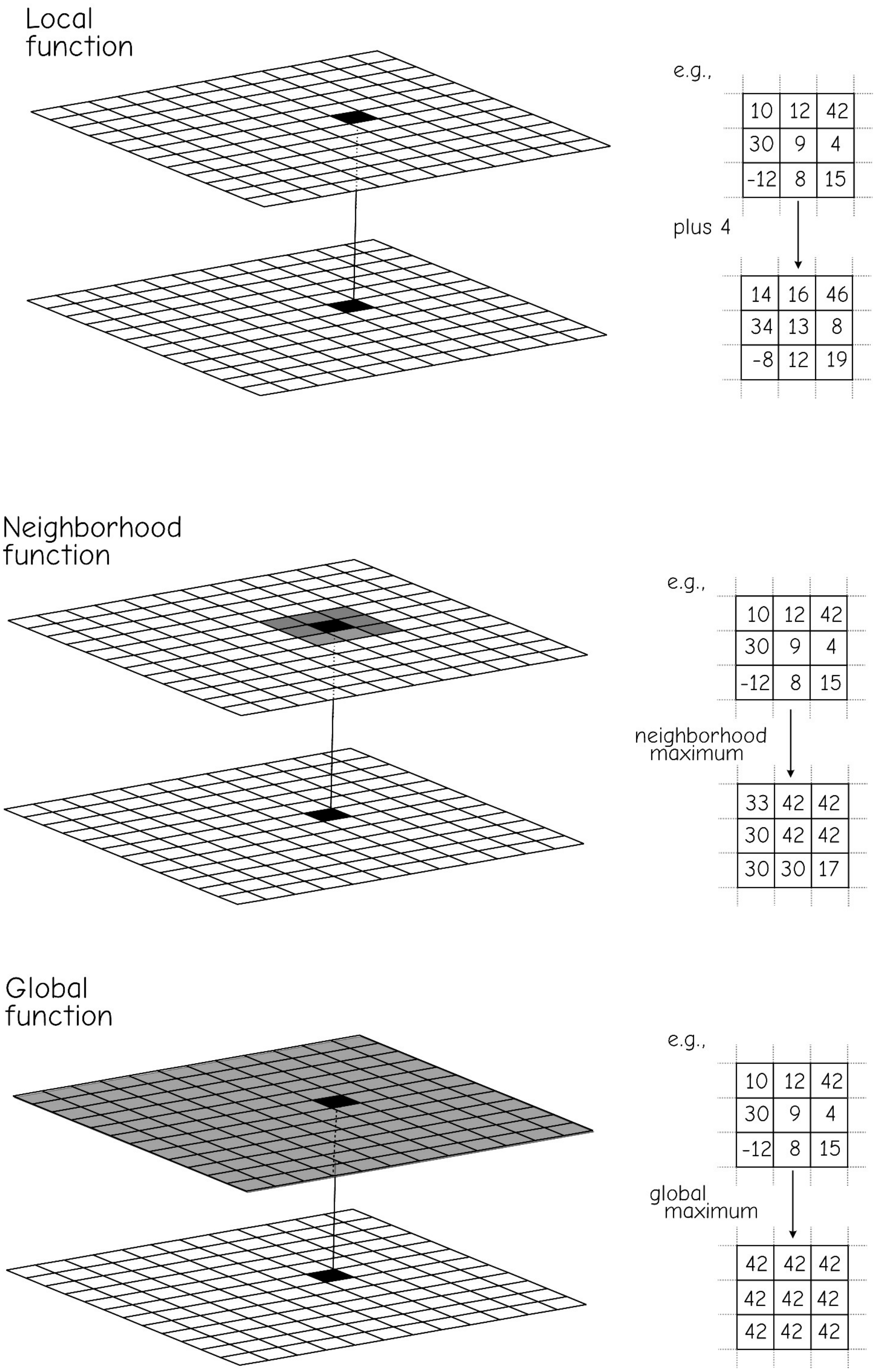


Figure 10-3: Raster operations may be local, neighborhood, or global. Target cells are shown in black, and source cells contributing to the target cell value in the output are shown in gray. Local operations show a cell-to-cell correspondence between input and output. Neighborhood operations include a group of nearby cells as input. Global operations use all cells in an input layer to determine values in an output layer.

Local Functions

There is a broad number of local functions (or operations) that can be conveniently placed in one of four classes: mathematical functions, Boolean or logical operations, reclassification, and multilayer overlay.

Mathematical Functions

We may generate a new data layer by applying mathematical functions on a cell-by-cell basis to input layers (Figure 10-1, Table 10-1). Any number of inputs and outputs may be supported, depending on the function.

A broad array of mathematical functions may be used, with a few constraints. Raster data value and type are perhaps the most common constraints. Most raster models store one data value in a cell. Each raster data set has a data type and maximum size that applies to each cell; for example, a two-byte signed integer may be stored. Mathematical operations that create noninteger values, or values larger than 32,768 (the capacity of a two-byte integer), may not be stored accurately in a two-byte integer output layer. Most systems will do some form of automatic type conversion, but there are often limits on the largest values that can be stored, even with automatic conversion.

Although the set of functions and function names differ among software packages, nearly all packages support the basic arithmetic operations of addition through division, and most provide the trigonometric functions and their inverses (e.g., sin, asin). Truncation, power, and modulus functions are also commonly supported, and vendors often include additional functions they perceive to be of special interest. These mathematical functions are often applied in raster analysis, for example, when multiplying each cell by 3.28 to convert height values from units in meters to units measured in feet.

Table 10-1: Common local mathematical functions.

Function	Description
Add, subtract, multiply, and divide	cell-by-cell combination with the arithmetic operation
ABS	Absolute value of each cell
EXP, EXP10, LN, LN10	Applies base e and base 10 exponentiation and logarithms
SIN, COS, TAN, ASIN, ACOS, ATAN	Apply trigonometric functions on a cell-by-cell basis
INT, TRUNC	Truncate cell values, output integer portion
MODULUS	Assigns the decimal portion of each cell
ROUND	Rounds a cell value up or down to nearest integer value
SQRT, ROOT	Calculates the square root or specifies other root of each cell value
POWER	Raises each cell to a defined power

Note that although many systems will let you perform these operations on any type of raster data, they often only make sense for interval/ratio data, and may return erroneous results when applied to nominal or ordinal data. Numbers may be assigned to indicate population density by high, medium, and low, and while the sin function may be applied to these data, the results will usually have little meaning.

Logical Operations

There are many local functions that apply logical (also known as Boolean) operations to raster data. A logical operation typically involves the comparison of a cell to a scalar value or set of values and outputs a "true" or a "false" value. True is often represented by an output value of 1, and false by an output value of 0.

There are three basic logical operations, AND, OR, and NOT (Figure 10-4). The AND and OR operations require two input layers. These layers serve as a basis for comparison and assignment. AND requires both input values be true for the assignment of a true output. Typically, any nonzero value is considered to be true, and zeros false. Note in Figure 10-4a that output values are typically restricted to 1 and 0, even though there may be a range of input values. Also note that there may be cells where no data are recorded. How these are assigned depends on the specific GIS system. Most systems assign null output when any input is null; others assign false values when any input is null.

Figure 10-4b shows an example of the OR operation. This cell-by-cell comparison assigns true to the output if either of the corresponding input cells is true. Note that the cells in either layer or both layers may be true for a true assignment, and that in this example, null values (N) are assigned when either of the inputs is null. Some implementations assign a true value to the output cell if any of the inputs is non-null and non-zero;

Input | Output

a)

1	3	1	1		0	1	0	9		0	1	0	1
0	N	2	-1	AND	0	5	2	5	=	0	N	1	1
1	2	5	0		0	2	N	2		0	1	N	0
0	1	N	N		0	-3	4	8		0	1	N	N

b)

1	3	1	1		0	1	0	9		1	1	1	1
0	N	2	-1	OR	0	5	2	5	=	0	N	1	1
1	2	5	0		0	2	N	2		1	1	N	1
0	1	N	N		0	-3	4	8		0	1	N	N

c)

	1	3	1	1		0	0	0	0
NOT	0	N	2	-1	=	1	N	0	0
	1	2	5	0		0	0	0	1
	0	1	N	N		1	0	N	N

Figure 10-4: Examples of logical operations applied to raster data. Operations place true (non-zero) or false values (0) depending on the input values. AND requires both corresponding input cells be true for a true output, OR assigns true if either input is true, and NOT simply reverses the true and false values. Null or unassigned cells are denoted with an N.

the reader should consult the manual for the specific software tool they use.

Figure 10-4c shows an example of the NOT operation. This operation switches true for false, and false for true. Note that null input assigns null output.

Finally, note that many systems provide an XOR operation, known as an eXclusive OR (not illustrated in our examples). This is similar to an OR operation, except that true values are assigned to the output when only one or the other of the inputs is true, but not when both inputs are true. This is a more restrictive case than the general OR, and may be used in instances when we wish to distinguish among origins for a true assignment.

Logical operations may be provided that perform ordinal or equality comparisons, or that test if cell values are null (Figure 10-5). Ordinal comparisons include less than, greater than, less than or equal to, greater than or equal to, equal, and not equal. Examples of these logical comparisons are shown in Figure 10-5a and b, respectively. These operations are applied cell-by-cell, and the corresponding true or false output assigned. As shown in Figure 10-5a, the upper left cell of the first input layer is not less than the upper left cell of the second input layer, so a 0 (false) is assigned to the upper left cell in the output layer. The upper right cell in the first layer is less than the corresponding cell in the second input layer, resulting in the assignment of 1 (true) in the output layer.

We often need to test for missing or unassigned values in a raster data layer. The operation has no standard name, and may be variously called via ISMISSING, ISNULL, or some other descriptive name. The operation tests each cell for a null value, shown as N in Figure 10-5c. A 0 is assigned to the corre-

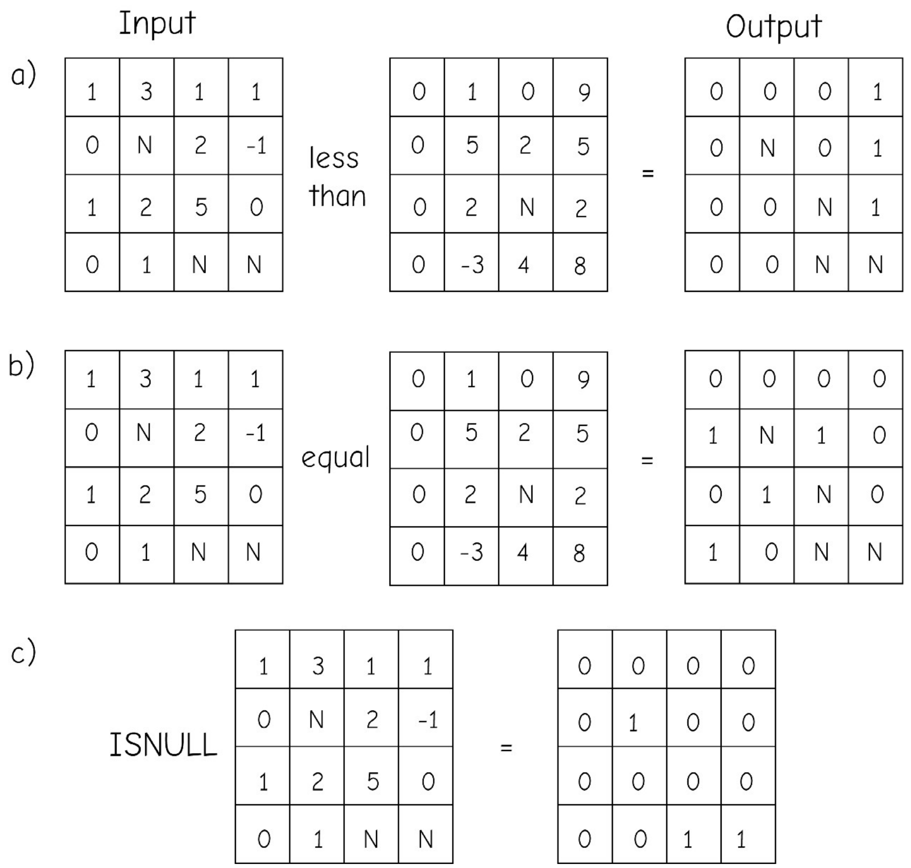


Figure 10-5: Examples of logical comparison operators for raster data sets. Output values record the ordinal comparisons for inputs, with a 1 signifying true and a 0 false. The ISNULL operation assigns true values to the output whenever there are missing input data.

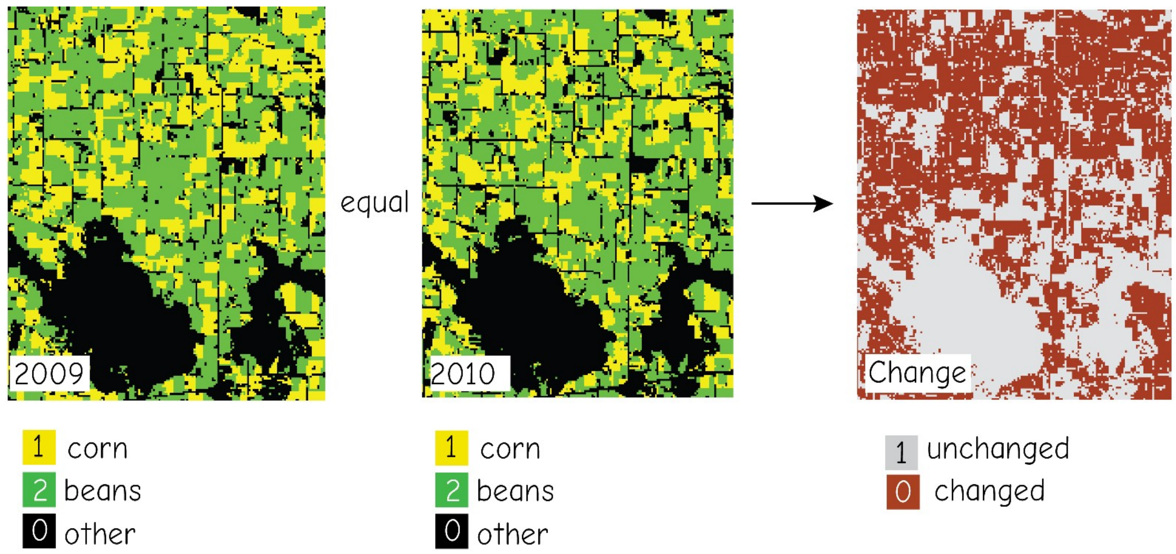


Figure 10-6: An example of a logical operation applied to categorical data, here landuse classes of corn (value = 1), soybeans (2), and other crops (0) over two time periods. The equal operation returns 0 where land use has changed, and 1 where it has remained constant over these two periods.

sponding output cell if a non-null value is found, otherwise a 1 is assigned. These tests for missing values are helpful when identifying and replacing missing data, or when determining the adequacy of a data set and identifying areas in need of additional sampling.

Figure 10-6 shows an example of a logical comparison among two data layers. The left and central panels show landcover for an agricultural area, with three categories: corn (1), soybeans (2), and all others (0). We may be interested in identifying acres that were rotated between these two crops, or from these two crops to other crops over the 2009-2010 time period. The logical equal comparison between these layers reveals areas that have changed. If the cell values are not equal across the years, the logical equal comparison will return a value of 0, while areas that remain the same will maintain the value of 1. Further logical comparisons, using class values, could identify how much each of the component crop types had changed.

Note that logical operators may be applied to interval/ratio, ordinal, and categorical data, although ordinal comparisons should be carefully applied to categorical data. For example, in our crop types example in Figure 10-6, soybeans are assigned a value of 2, and are "larger" than corn, but this distinction does not imply that soybeans are somehow two times larger, more valuable or anything other than just different from corn.

Reclassification

Raster reclassification assigns output values that depend on the specific set of input values. Assignment is most often defined by a table, ranges of values, or a conditional test.

Raster reclassification by a table is based on matching input cell values to a reclassification table (Figure 10-7a). The reclassification table specifies the mapping between input values and output values. Each input cell value is compared to entries for an "in" column in the table. When a match is found, the corresponding "out" value is assigned to the output raster layer. Unmatched input values can be handled in one of several ways. The most logically consistent manner is to assign a null value, as shown in Figure 10-7a for the input value of -1. Some software simply assigns the input cell value when there is no match. As with all spatial processing tools, the specifics of

the implementation must be documented and understood.

Figure 10-7b illustrates a reclassification by a range of values. This process is similar to a reclassification by a table, except that a range of values appears for each entry in the reclassification table. Each range corresponds to an output value. This allows a more compact representation of the reclassification. A reclassification over a range is also a simple way to apply the automated reclassification rules discussed at length in Chapter 8 — the equal interval, equal area, natural breaks, or other automated class-creation methods. These automated assignment methods are often used for raster data sets because of the large number of values they contain.

Data can also be reclassified to select the input source based on a condition. These "conditional" functions have varying syntax, but typically require a condition that results in a true or false outcome. The value or source layer assigned for a true outcome is specified, as is the value or source layer assigned for a false outcome. An example of one conditional function may be:

Output = CON (test, out if true, out if false) (10.4)

where CON is the conditional function, test is the condition to be tested, out if true defines the value assigned if the condition tests true, and out if false defines the value assigned if the condition tests false (Figure 10-8). Note that the value that is output may be a scalar value, for example, the number 2, or the value output may come from the corresponding location in a specified raster layer. The condition is applied on a cell-by-cell basis, and the output value assigned based on the results of the conditional test.

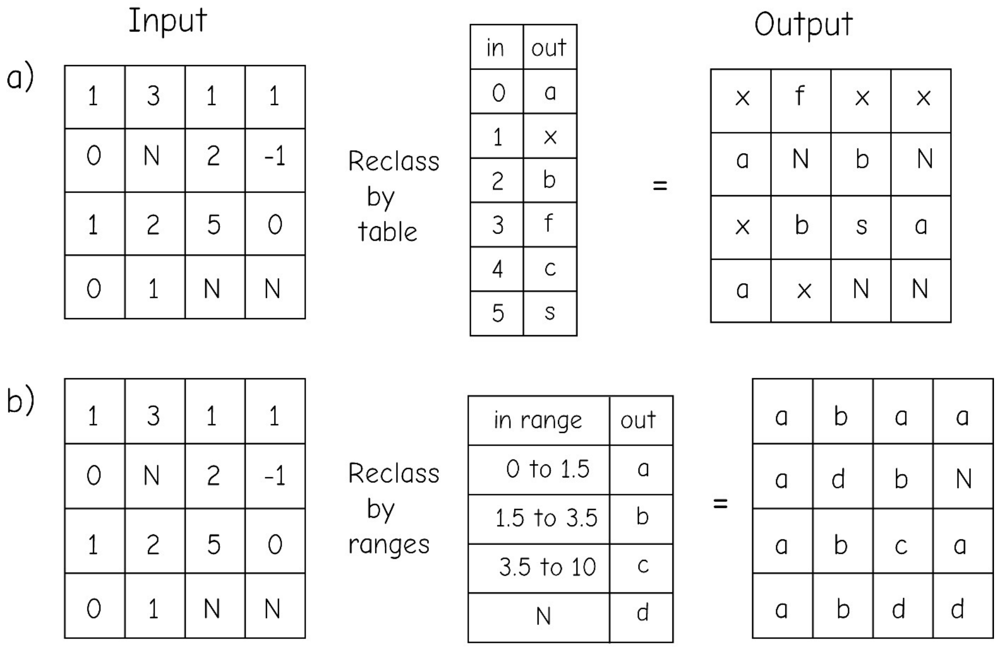


Figure 10-7: Raster reclassification by table matching (a) and by table range (b). In both cases, input cell values are compared to the "in" column of the table. A match is found and the corresponding "out" values assigned.

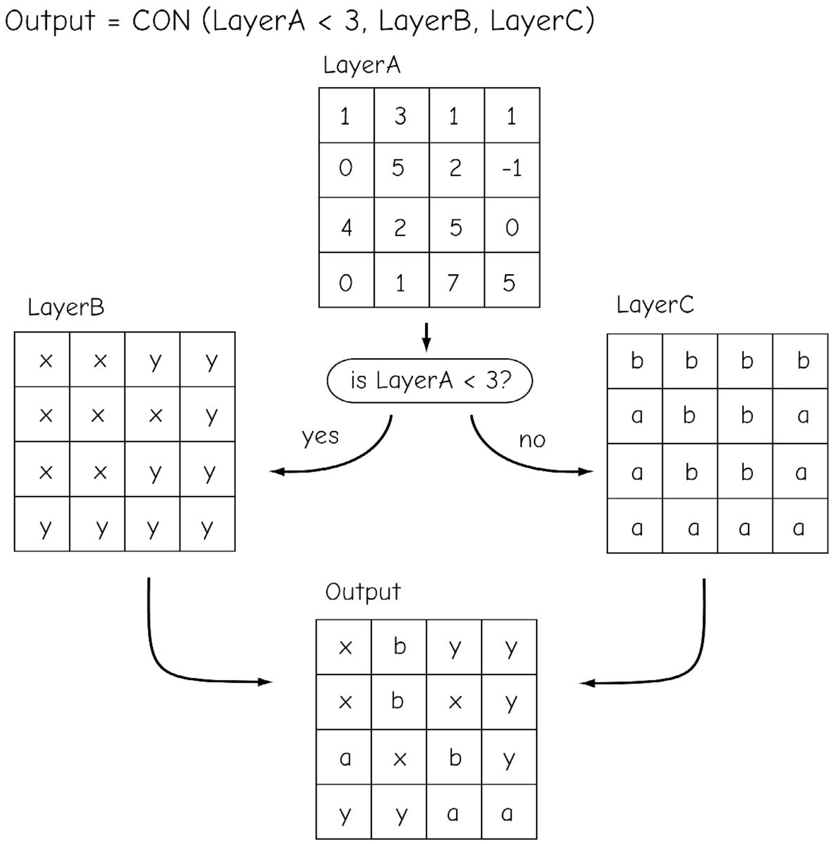


Figure 10-8: Reclassification by condition assigns an output based on a conditional test. In this example, each cell in LayerA is compared to the number 3. For LayerA cells with values less than 3, the condition evaluates to be true, and the output cell value is assigned from LayerB. If the LayerA cell value is equal to or greater than 3, then the output cell value is assigned from LayerC.

Nested Functions

Local functions may be nested in analyses. Functions are nested when a function is used as the argument of another function. For example, we may wish to take the natural logarithm (LN) of all the cells in a layer. The mathematical LN function is only defined for positive values. When inputs are negative, we need to either accept null values in the output data layer, or process these input cells in a different manner. We could do this by applying the absolute value function (ABS) to create an intermediate output. We could then apply the LN function to this output for our final result. This could be described as the equations:

InitOutput = ABS (Input_Layer) (10.5)

FinalOutput = LN (InitOutput). (10.6)

We could do the same thing by nesting the functions, if allowed by the GIS software:

FinalOutput = LN (ABS (Input_Layer)) (10.7)

Figure 10-9 shows another example of nested function. Output values are assigned from two different input layers. Cell values are assigned from LayerB when LayerA values are null, and from LayerC when LayerA values are non-null. This might be desirable if we have an incomplete but otherwise high-quality data set, and we wish to fill missing values from the next best available data. Map algebraic expressions with nested functions can become quite complex, but also may be quite effective and efficient in solving complex problems.

Overlay

Raster overlay combines features from two or more data layers, and is among the most useful of spatial functions. The features in raster data correspond to cells, or perhaps groups of cells with the same values, but as with vector overlay, great utility is often gained from combining data from different layers.

There are some differences between raster and vector overlay due to the differences in the data model. Raster overlay is often restricted to nominal data. The cell values do not typically represent continuous variables such as temperature, but rather categorical variables such as type or township name.

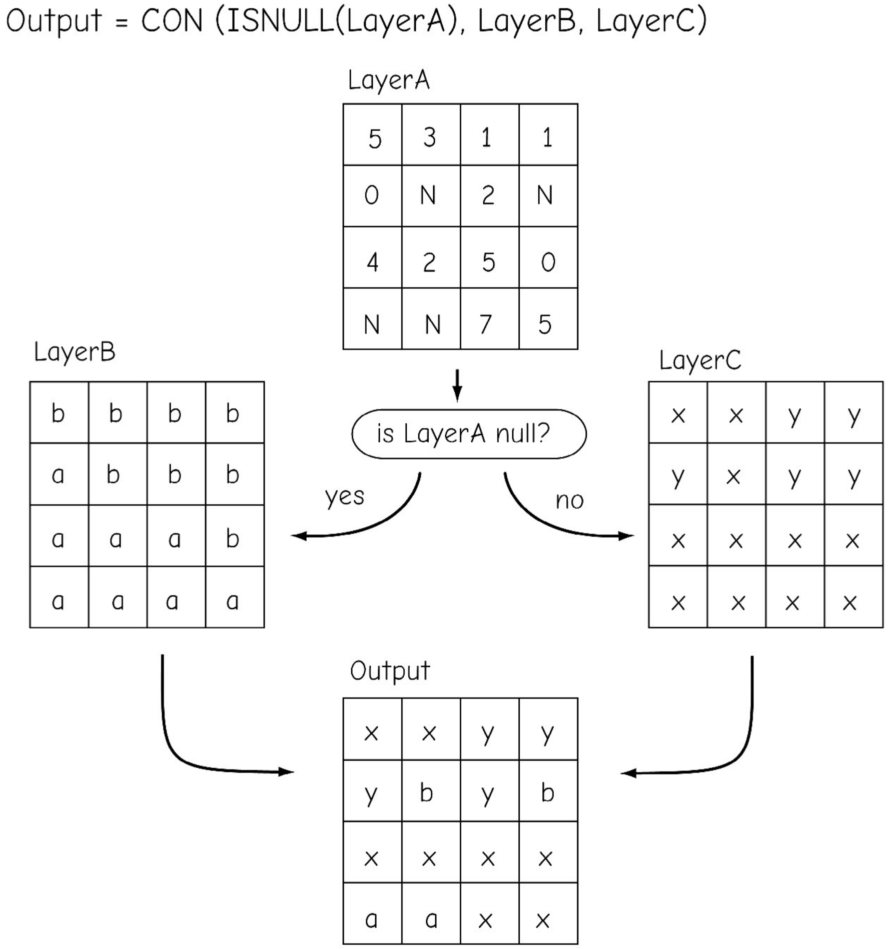


Figure 10-9: Local logical operations may be combined. This example shows the ISNULL function embedded in the conditional function, CON. As described in Figure 10-7, the first argument in this CON function defines the Boolean test. Here, if the cell value in LayerA is null (N), then the conditional test returns a true value. This executes the following entry, an assignment to the output from LayerB. If the cell is non-null, ISNULL returns a value of false, and the Output cell value assigned from the corresponding location in LayerC. Note the Output values are a or b, from LayerB, for the cells in LayerA that have null (N) values.

Although raster overlay may be implemented so that it admits continuous numbers, this typically results in too many unique cell combinations to be of much value. If continuous data are used, they are often converted to categories first; for example, rainfall may be assigned to low, medium, or high classes.

Raster overlay involves the cell-by-cell comparison of values in two or more layers (Figure 10-10). The values in each input data layer are associated with a specific combination of additional variables, and these additional variables may be recorded in an attribute table. Each unique combination of cells from the two layers is identified, and assigned a new identifier (Out-ID) in the Output layer. Note the two input attribute tables are combined in a corresponding fashion. In Figure 10-10 you can see the upper left corner of the Output layer has the corresponding type and name attribute values from Input layer 1, and the ID and cost attribute values from Input layer 2.

Recall that in many implementations of the raster data model there is a many-to-one relationship between the raster cells and the attribute rows. This occurs because multiple cells correspond to each row. Also note that the cells may form disjunct regions of the

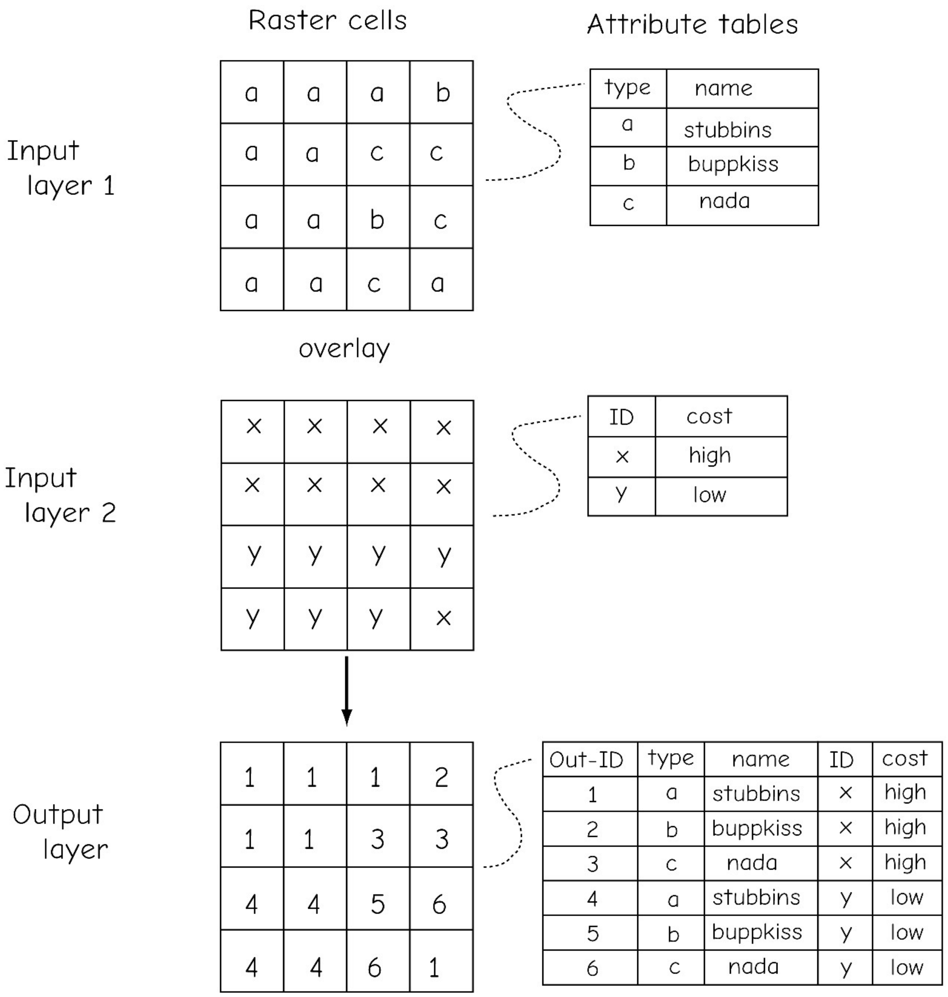


Figure 10-10: Raster overlay involves the cell-by-cell combination of data in multiple layers. New output values are generated for each unique combination of input values.

same type; for example, Figure 10-10 shows a cell of type a in the lower left corner of Input layer 1 that is not contiguous with the rest of the a cells in Layer 1. This combination carries through to the output, where there are disjunct groups of cells with Out-ID values of 6.

Clip (or extraction) is another common type of local raster function (Figure 10-11). Source and template data layers are specified and an output data layer created. This output layer contains only the values of the source that are indicated by the template layer. The nature of the template and output data layer values depend on the specific implementation of the raster extraction. Template values are typically assigned a value of 1 for those cells that are to pass through to the output, and a 0 or null value for those that are to be ignored. Output values for the clipped area are copied from the source, while output values for the area outside the clipped region are typically assigned a null value, or the value 0.

Care must be taken to ensure there are no ambiguous cells created by this convention. For example, if there are null values in both the source data layer and the area outside the clip, one cannot be certain if the nulls come from the source or indicate a region outside the clip area. Special coding or other provisions can be used to avoid these ambiguities.

Overlay functions in map algebra can be created through addition and multiplication functions. Union operations can be performed with layer addition, and clip operations through multiplication. These two raster algebra operations may be used to combine raster data layers in a number of ways, even if the specific software implementation of the raster GIS does not provide an explicit or specialized overlay function.

The union of two raster data layers may be performed through addition. When two layers are added, values are added on a cell-by-cell basis and the results placed in corresponding cell locations for an output data layer. Each output value may be used to identify the combinations of input values.

Figure 10-12 shows the overlay of two data layers through raster addition. Cells in Layer A have values 10 through 40, and cells in Layer B have values 1 through 7. These might correspond to four different species types in Layer A and six different landuse types in Layer B. Data layers are combined on a cell-by-cell basis, so each cell value in Layer A is added to the corresponding cell value in Layer B. In the upper right corner cell, the value 20 from Layer A is added to 2 in Layer B, and the resultant value 22 is placed in the Output layer. Cell addition of these two layers will result in a set of numbers between 12 and 47. These numbers correspond to the various combinations of species and landuse.

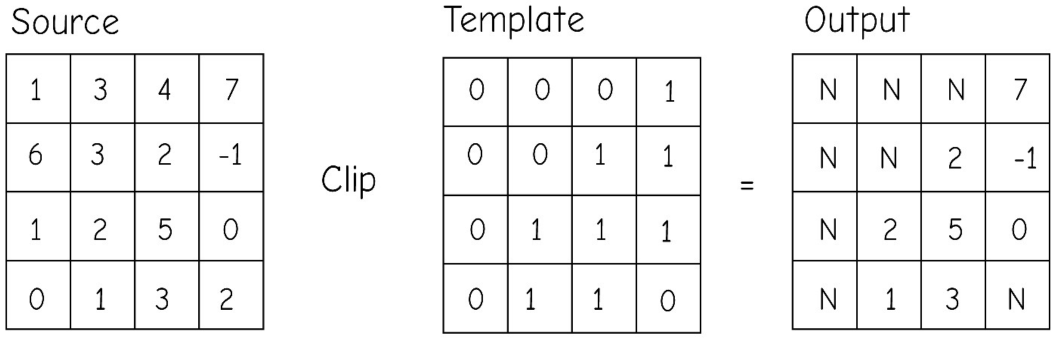


Figure 10-11: An illustration of a raster clip (or extraction) operation. Values from a source layer are extracted based on values in an extraction template. Output cells are often assigned a null value, N, in the "outside" area.

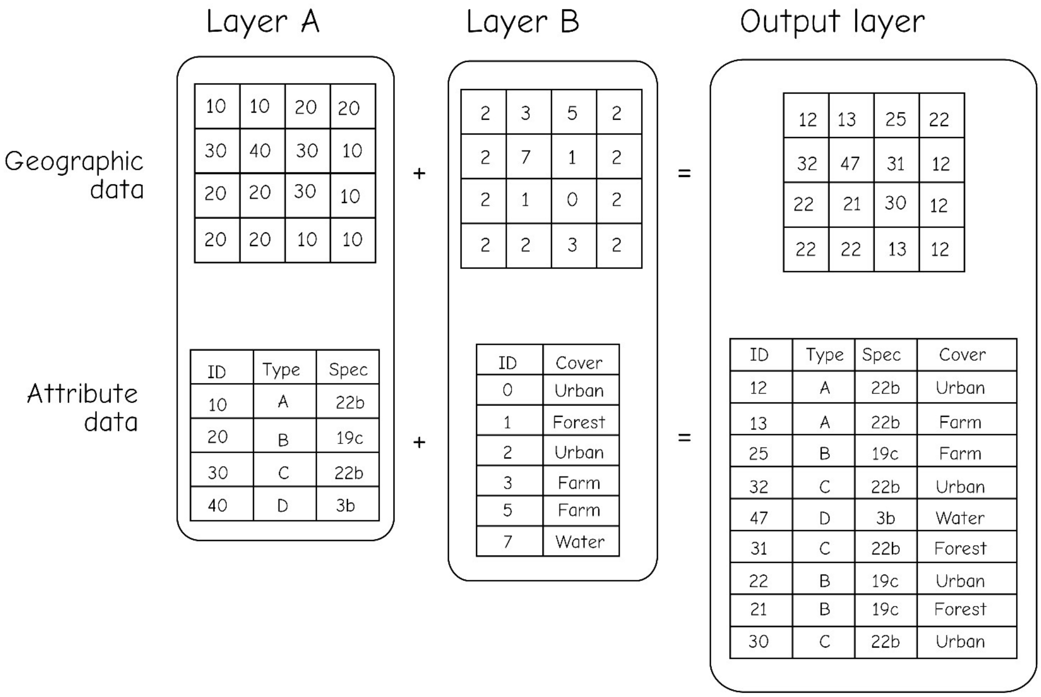


Figure 10-12: Overlay through raster addition. Cell values in Layer A and Layer B are added to yield values in the Output layer. Attribute values associated with each input layer may be combined in an associated table.

Attribute data may be matched to the corresponding attribute tables. Each unique value in the input data layer is identified and associated with the appropriate combinations in the output data layer. Attribute assignment is illustrated in Figure 10-12. Cells in Layer A with an ID value of 10 are associated with Type = A and Spec = 22b. Every cell in the Output layer that exhibit an ID = 10 for input Layer A have a Type = A and Spec = 22b in the corresponding output cells.

Note that identical values may be discontinuous in the output data layer. The ID values of 12 are found in two disjunct sets in Figure 10-12, one in the upper left corner and one in the lower right corner of the output table. These cells with ID=12 are all referred to by the same entry in the attribute table. This many-to-one relationship may occur quite often; for example, it also exists for ID=22 in the Output layer.

Raster data sets often do not uniquely identify disjunct but otherwise identical areas because of limits on attribute table size. Raster data sets often have thousands to millions of unique cell locations. The average area of contiguous, identical cells typically decreases when data layers are combined through raster overlay. There are often many small areas that have the same combination of input values, but are separated from other cells with the same value. If each group of cells is assigned a unique identifier, the attribute table may grow to be quite large. Large data sets are becoming less of a handicap as computing power and space increases, but many software packages by default implement a one-to-many relationship. This contrasts with the common approach applied by vector overlay software packages, which typically identify each polygon uniquely, whether or not there are other polygons with an identical set of attribute values. The GIS user needs to understand the output convention for the specific software so that output from overlay opera-

tions may be properly interpreted and applied.

Raster addition can be used to mimic raster overlay. However, care must be taken to avoid ambiguous combinations when using raster layer addition for overlay. Identical output numbers derived from two different combinations must be avoided, because it represents an ambiguous result. Consider the example shown in Figure 10-13. Two data layers are overlaid through addition. A value of 4 may occur in the output layer from multiple combinations: 2 in Layer A and 2 in Layer B, or 1 in Layer A and 3 in Layer B. There are similar problems for output values equal to 3, so our results are ambiguous. We must ensure this cannot occur. We typically do this by reclassifying a data layer. For example, we could multiply Layer A by 10, thus giving values of 10, 20, 30, and 40. The output values will then uniquely identify the combination of inputs.

A clip using raster data may be implemented as a reclassification and then a multiplication. Note that in Chapter 9, we described the vector clip function as a special case of overlay in which the attributes and interior geometry were saved based on the boundaries in a clipping layer. This clipping layer serves as an outline or area template for which data are retained. In a raster clip, the clipping layer may be represented as a set of cells with a value of 1 embedded in nonclipping cells with values of 0.

Figure 10-14 illustrates a raster clip operation that is a combination of cell reclassification and multiplication. The first step is to identify the set of values that defines the clip area. This is the portion of the input data layer to be transferred to the output data layer. Individual cell values or cell values over an interval or range may be defined. These may come from a selection based on raster values, from a list of values, or from a previous spatial operation such as a buffer.

A clip template is created that defines a *binary mask*, a set of cells that "mask" out a portion of an input layer. Cells to be passed through to the output layer are set to the value 1 (Figure 10-14). Cells to be "clipped away" are set to the value 0. The clip template or layer is then multiplied by the input

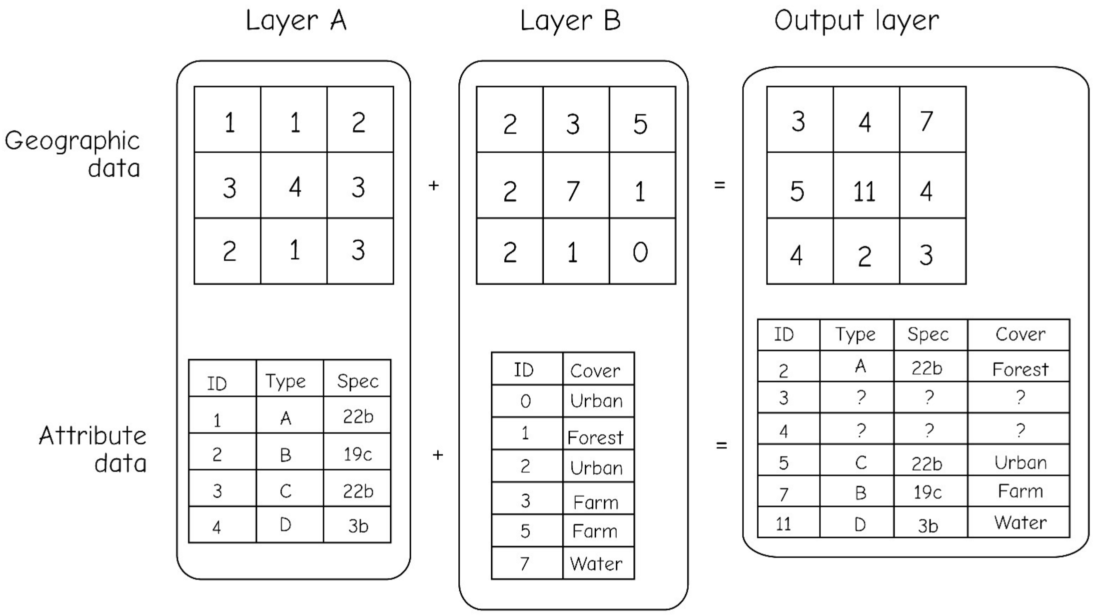


Figure 10-13: Raster addition will lead to ambiguous output when different combinations of inputs lead to identical output combinations (unknown type, spec, and cover in the output layer). Input layers may be classified or renumbered to ensure unique output combinations.

raster, yielding an output raster. Cell-by-cell multiplication by 1 passes values through to the output layer. Multiplication by 0 discards values for the cell, resulting in a clipped raster to the area of interest.

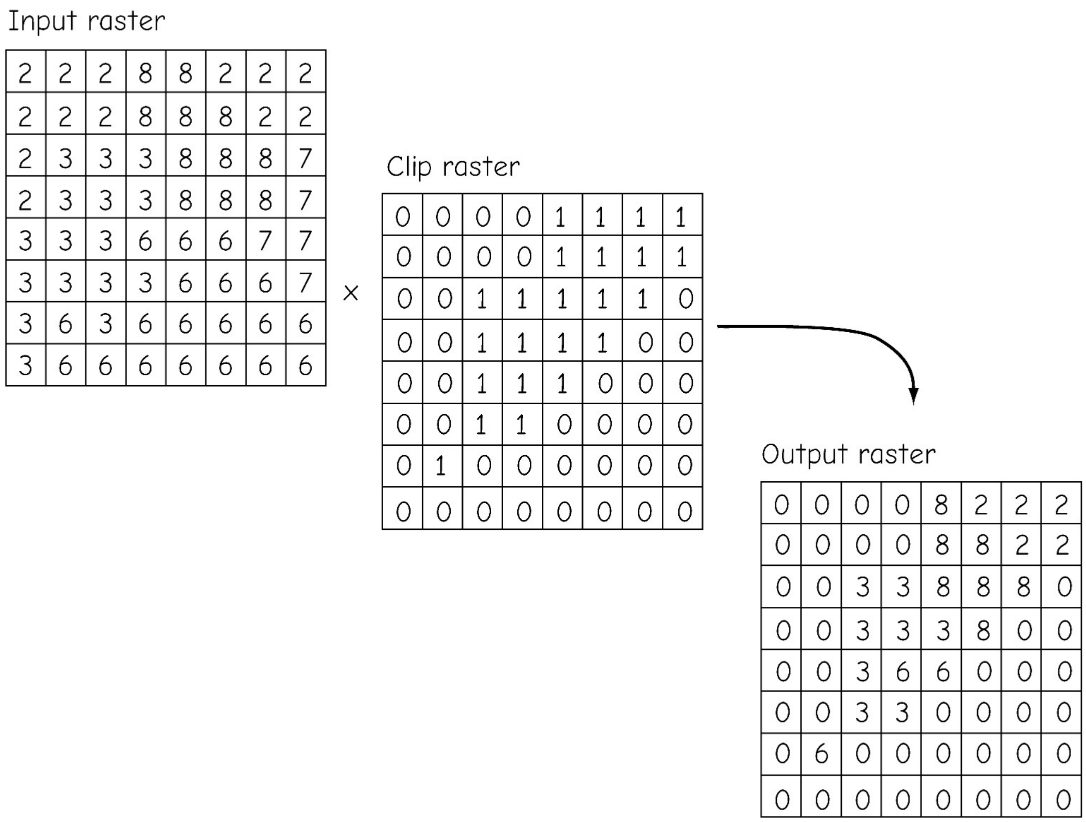


Figure 10-14: A raster clip operation may be performed via multiplication of a binary mask. An input raster (left) has been multiplied by a binary raster. Cell-by-cell multiplication passes through each input cell corresponding to a 1 in the clip raster (center), and passes through a 0 for all other cells to the output raster (right). Note that zero values are not allowed in the input raster because they have the same effect as zeros in the clip raster.

Neighborhood, Zonal, Distance, and Global Functions

Neighborhood functions (or operations) in raster analyses deserve an extended discussion because they offer substantial analytical power and flexibility. Neighborhood operations are applied in many analyses across a broad range of topics, including the calculation of slope, aspect, and spatial correlation.

Neighborhood operations most often depend on the concept of a *moving window*. A "window" is a configuration of raster cells used to specify the input values for an operation (Figure 10-15). The window is positioned on a given location over the input raster, and an operation applied that involves the cells contained in the window. The result of the operation is usually associated with the cell at the center of the window position. The result of the operation is saved to an output layer at the center cell location. The window is then "moved" to be centered over the adjacent cell and the computation repeated (Figure 10-15). The window is swept across a raster data layer, usually from left to right in successive rows from top to bottom. At each window location, the moving window function is calculated and the result output to the new data layer.

Moving windows are defined in part by their dimensions. For example, a 3 by 3 moving window has an edge length of three cells in the x and y directions, for a total area of nine cells. Moving windows may be any size and shape, but they are typically odd-

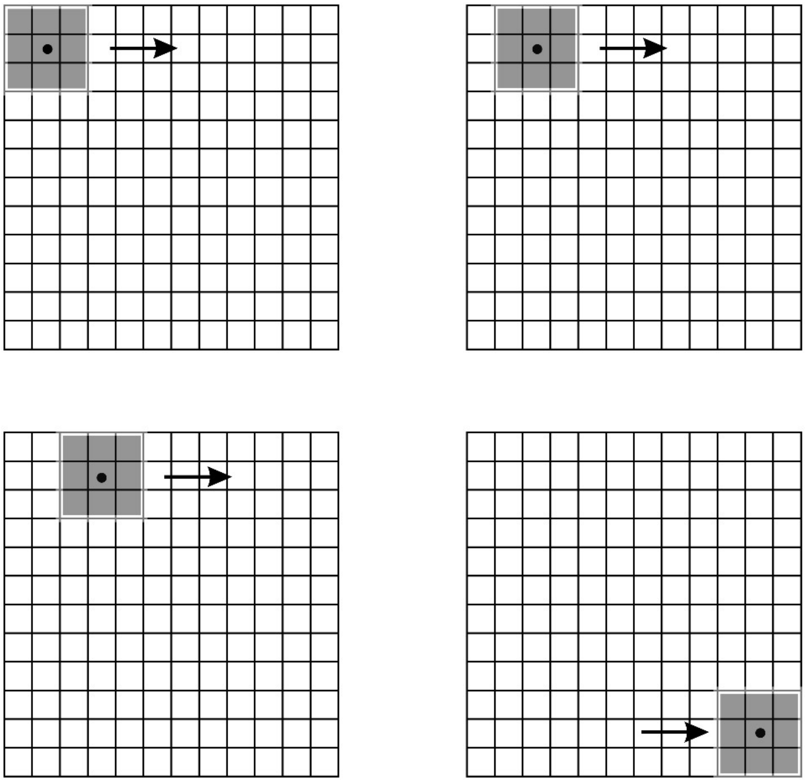


Figure 10-15: The concept of a moving window in raster neighborhood operations. Here, a 3 by 3 window is swept from left to right and from top to bottom across a raster layer. The window at each location defines the input cells used in a raster operation.

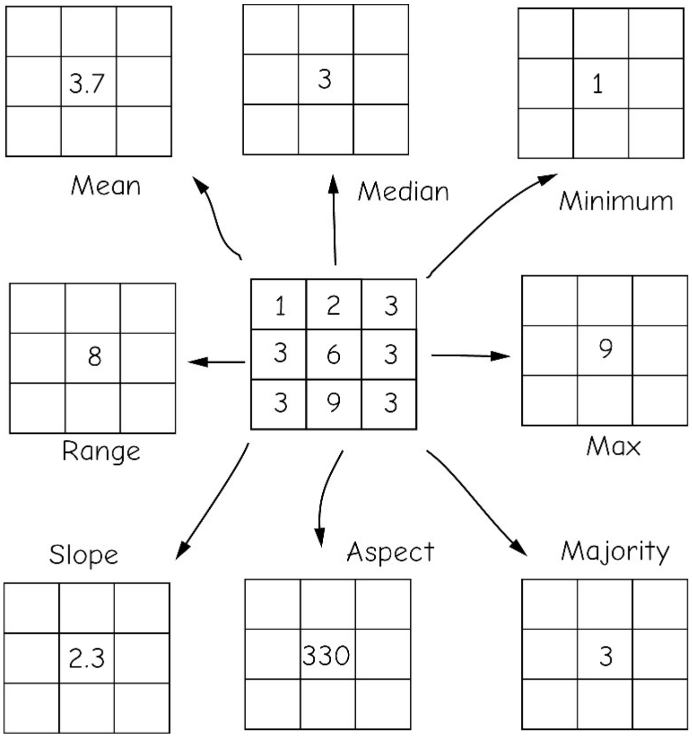


Figure 10-16: A given raster neighborhood may define the input for several raster neighborhood operations. Here a 3 by 3 neighborhood is specified. These nine cells may be used as input for mean, median, minimum, or a number of other functions.

numbered in both the x and y directions to provide a natural center cell, and they are typically square. A 3 by 3 cell window is the most common size, although windows may also be rectangular. Windows may also have irregular shapes; for example, L-shaped, circular, or wedge-shaped moving windows are sometimes specified.

There are many neighborhood functions that use a moving window. These include simple functions such as mean, maximum, minimum, or range (Figure 10-16). Neighborhood functions may be complicated, for example, the statistical standard deviation, and they may be nonarithmetic, as the functions that return a count of the number of unique values, or the mode, or a Boolean occurrence. Any function that combines information from a consistently shaped group of raster cells may be implemented with a moving window.

Moving window functions may be arithmetic, adding, subtracting, averaging, or otherwise mathematically combining the values around a central cell, or they may be comparative or otherwise extract values from a set of cells. Common statistical operations include calculating the largest value, the mode (peak of a histogram), median (middle value), the range (largest minus smallest), or diversity (number of different values). These neighborhood operations are useful for many kinds of processing.

Consider the *majority* operation, also known as a *majority filter.* You might wonder why one would want to calculate a majority filter for a data layer. Data smoothing is a common application. We described in Chapter 6 how multiband satellite data are often converted from raw image data to landcover classification maps.These classifiers often assign values on a pixel basis, and often result in many single pixels of one landcover type embedded within another landcover type. These single pixels are often smaller than the minimum mapping unit, the smallest uniform area that we care to map. A majority filter is often used to remove this classification "noise."

A majority filter is illustrated in Figure 10-17. It illustrates NASS crop data for an area in central Indiana, based on classified satellite images. There are over 40 common landcover types in the area, but these have

been reclassified into the dominant types of developed (road), corn, beans, and other crops. Each pixel is 30 m across, and the image on the left is NASS data as delivered. Note that corn and beans dominate, but there are many "stray" pixels of a dissonant vegetation type embedded or on the edge of a dominant type in an area; for example, bean pixels in a corn field, or corn pixels in a bean field. These stray pixels usually do not represent reality, in that although there may be the isolated plant or two from previously deposited seed in these annual crop rotations, the patches almost never approach 30 meters in size. The embedded cells are most often mis-classifications due to canopy thinning or perhaps weeds below the crop, and are often below the minimum mapping unit.

The illustrated majority filter counts the values of the four cells sharing an edge with any given cell. If a majority, meaning three or more cells, are of a type, then the cell is output as this majority type (top, Figure 10-17). If a majority is not reached, as when only one or two cells add up to the most frequent type in the four bordering cells, then the center cell value is unchanged in the output (bottom, Figure 10-17). The removal of most of the single pixel "noise" by the majority filter can be observed in the classified image on the right side of Figure 10-17.

There may be many variants for a given operation. The majority filter just discussed may assign an output value if only two of the four adjacent cell values are most frequent, or use the 8 or 24 nearest cells to calculate a majority. The dependence of output on algorithm specifics should always be recognized when applying any raster operation.

Figure 10-18 shows an example of a mean calculation using a moving window. The function scans all nine values in the window. It sums them and divides the sum by the number of cells in the window, thus calculating the mean cell value for the input window. The multiplication may be repre-

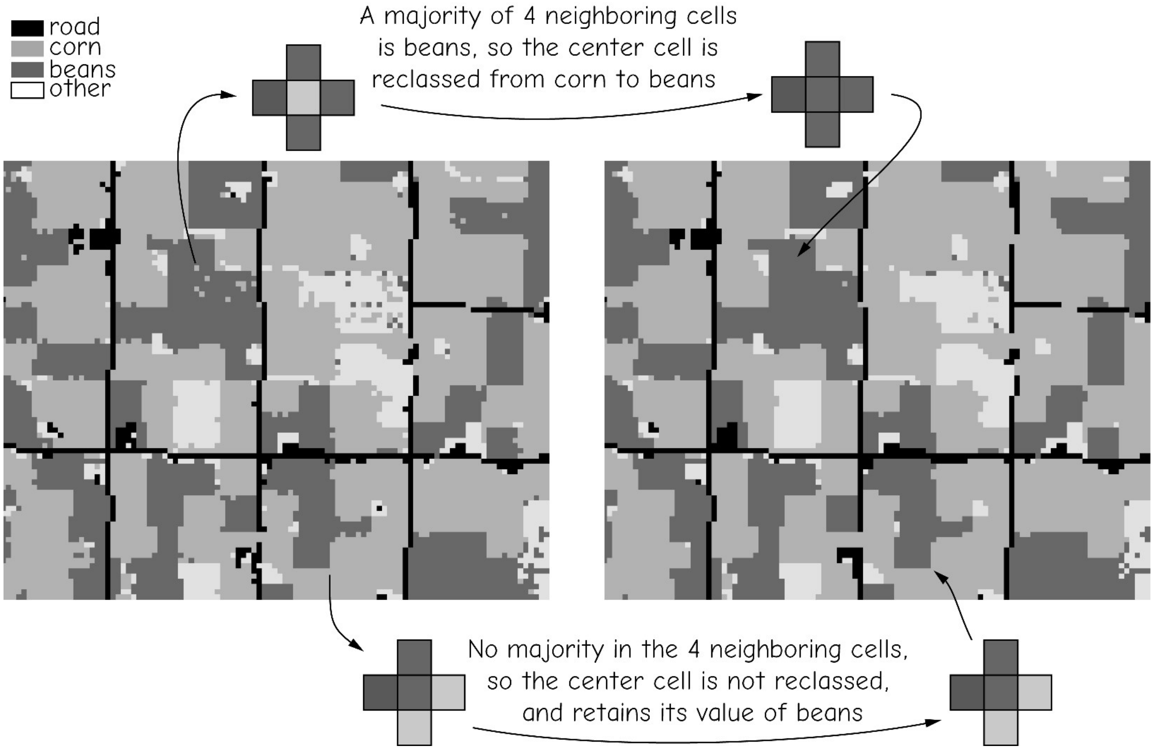


Figure 10-17: An example of a majority filter applied to a raster data layer from a classified satellite image. Many isolated cells are converted to the category of the dominant surrounding class.

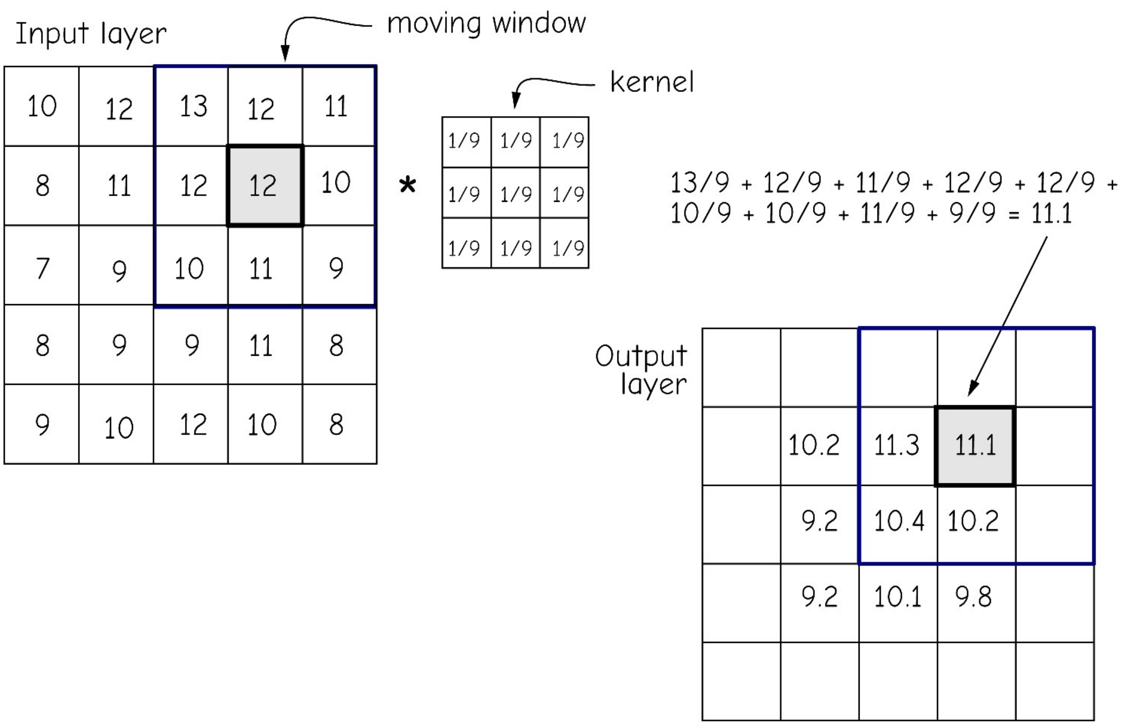


Figure 10-18: An example of a mean function applied via a moving window. Input layer cell values from a 3 by 3 moving window (upper left) are multiplied by a kernel to specify corresponding output cell values. This process is repeated for each cell in the input layer.

sented by a 3 by 3 grid containing the value one-ninth (1/9). The mean value is then stored in an output data layer in the location corresponding to the center cell of the moving window. The window is then shifted to the right and the process repeated. When the end of a row is reached the window is returned to the leftmost columns, shifted down one row, and the process repeated until all rows have been included.

The moving window for many simple mathematical functions may be defined by a *kernel*. A kernel for a moving window function is the set of cell constants for a given window size and shape. These constants are used in a function at every moving window location. The kernel in Figure 10-18 specifies a mean. As the figure shows, each cell value for the Input layer at a given window position is multiplied by the corresponding kernel constant. The result is placed in the Output layer.

Note that when the edge of the moving window is placed on the margin of the original raster grid, we are left with at least one border row or column for which output values are undefined. This is illustrated in Figure 10-18, right. The moving window is shown in the upper right corner of the input raster. The window is as near the top and to the right side of the raster as can be without placing input cell locations in the undefined region, outside the boundaries of the raster layer. The center cell for the window is one cell to the left and one cell down from the corner of the input raster. Output values are not defined for the cells along the top, bottom, and side margins of the output raster when using a 3 by 3 window, because a portion of the moving window would lie outside the raster data set. Each operation applied to successive output layers may erode the margin further.

There are several common methods of addressing this margin erosion. One is to define a study area larger than the area of interest. Data may be lost at the margins, but these data are not important if they are outside the area of interest. A second common approach defines a different kernel for margin windows (Figure 10-19). Margin kernels are similar to the main kernels, but modified as needed by the change in shape and size. Figure 10-19 illustrates a 3 by 3 kernel for the bulk of the raster. Corner values may be determined using a 2 by 2 kernel, and edges can be determined with 2 by 3 kernels. Outputs from these kernels are placed in the appropriate edge cells.

Different moving windows and kernels may be specified to implement many different neighborhood functions. For example, kernels may be used to detect edges within a raster layer. We might be interested in the difference in a variable across a landscape. For example, a railway accident may have caused a chemical spill and seepage through adjacent soils. We may wish to identify the boundary of the spill from a set of soil samples. Suppose there is a soil property, such as a chemical signature, that has a high concentration where the spill occurred, but a low concentration in other areas. We may apply kernels to identify where abrupt changes in the levels of this chemical create edges.

Edge detection is based on comparing differences across a kernel. The values on one side of the kernel are subtracted from the values on the other side. Large differences result in large output values, while small differences result in small output values. Edges are defined as those cells with output values larger than some threshold.

Figure 10-20 illustrates the application of an edge detection operation. The kernel on the left side of Figure 10-20 amplifies differences in the × direction. The values in the left of three adjacent columns are subtracted from the value in the corresponding right-hand row of cells. This process is repeated

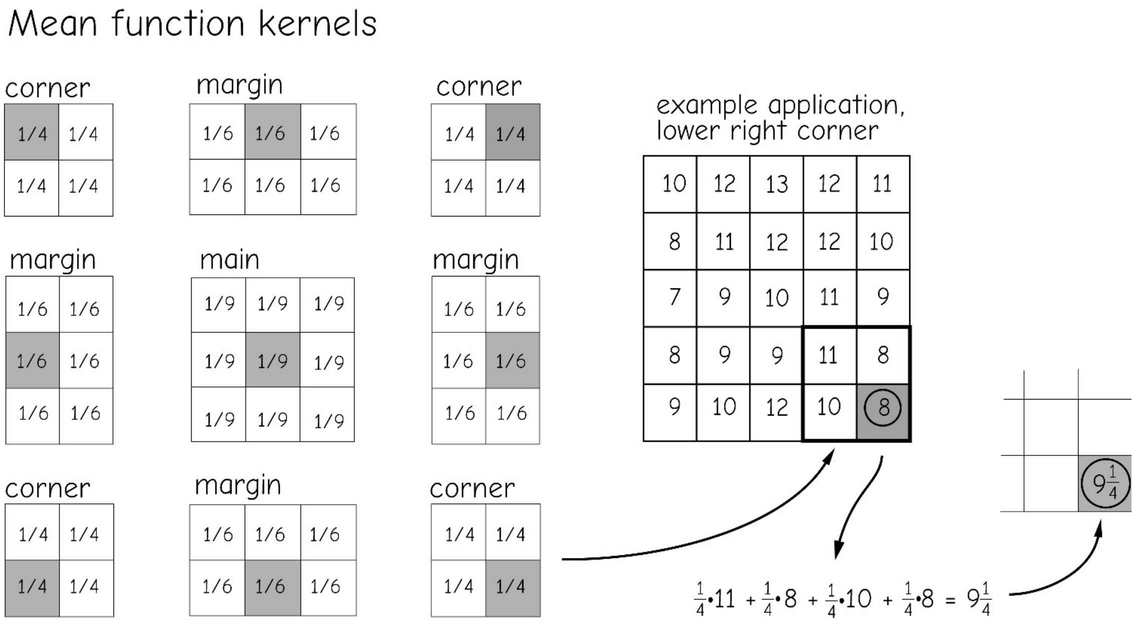


Figure 10-19: Kernels may be modified to fill values for a moving window function at the margin of a raster data set. Here the margin and corner kernels are defined differently than the "main" kernel. Output values are placed in the shaded cell for each kernel. The margin and corner kernels are similar to the main kernel, but are adjusted in shape and value to ignore areas "outside" the raster.

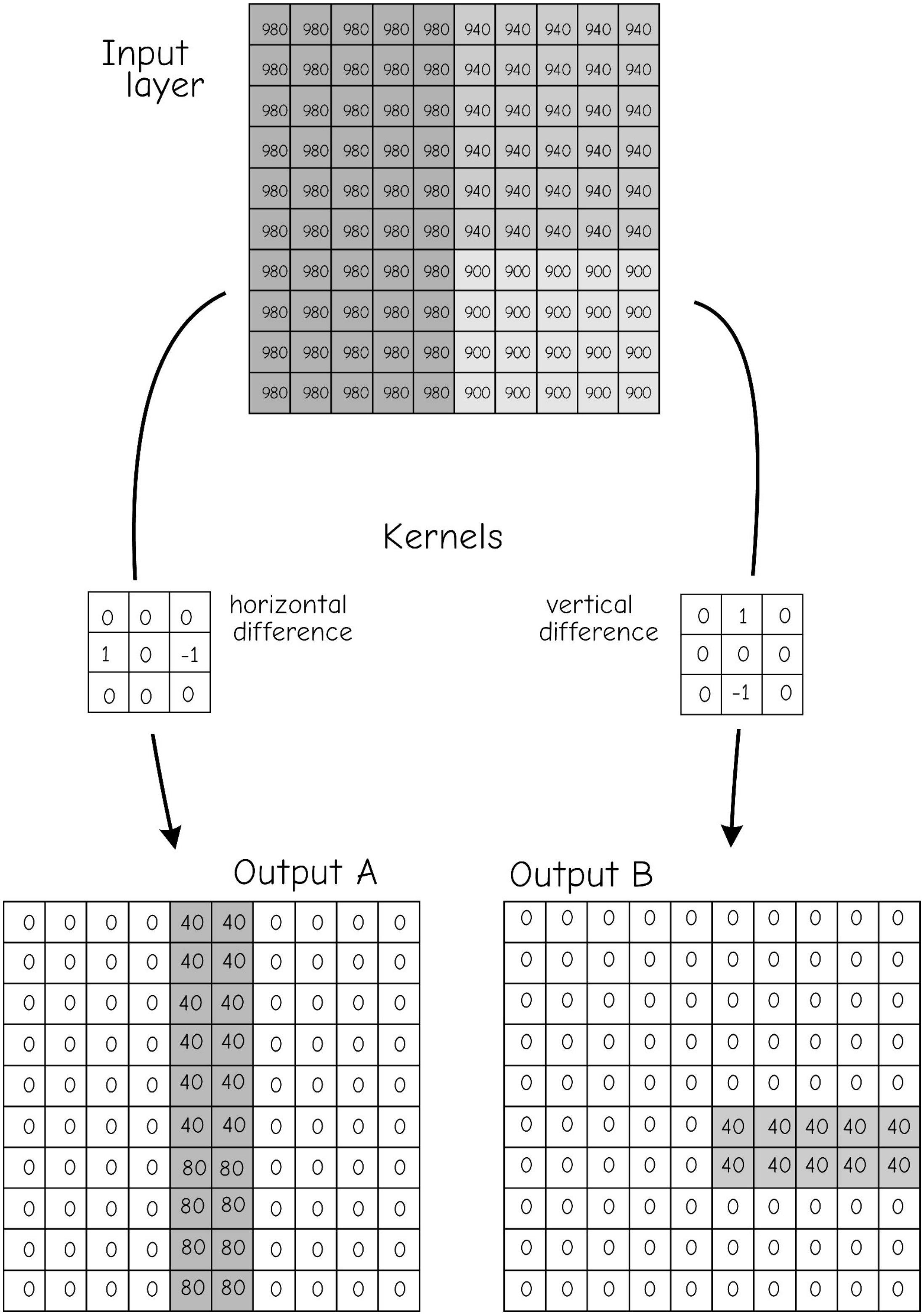


Figure 10-20: There is a large number of kernels used with moving windows. The kernel on the left amplifies differences in the x direction, while the kernel on the right amplifies differences in the y direction. These and other kernels may be used to detect specific features in a data layer.

for each cell in the kernel, and the values summed across all nine cells. Large differences result in large values, either positive or negative, saved in the center cell. Small differences between the left and right columns lead to a small number in the center cell. Thus, if there are large differences between values when moving in the x direction, this difference is highlighted. Spatial structure such as an abrupt change in elevation may be detected by this kernel. The kernel in the middle-right of Figure 10-20 may be used to detect differences in the y direction.

Neighborhood functions may also smooth the data. An averaging kernel, described above in Figure 10-18 and Figure 10-19, will reduce the difference between a cell and surrounding cells. This is because windows average across a group of cells, so there is much similarity in the output values calculated from adjacent window placements.

Raster data may contain "noise." Noise are values that are large or small relative to their spatial context. Noise may come from several sources, including measurement errors, mistakes in recording the original data, miscalculations, or data loss. There is often a need to correct these errors. If it is impossible or expensive to revisit the study area and collect new data, the noisy data may be smoothed using a kernel and moving window.

There are functions known as *high-pass filters* with kernels that accentuate differences between adjacent cells. These high-pass filter kernels may be useful in identifying the spikes or pits that are characteristic of noisy data. Cells identified as spikes or pits may then be evaluated and edited as appropriate, removing the erroneous values. High-pass kernels generally contain both negative and positive values in a pattern that accentuates local differences.

Figure 10-21 demonstrates the use of a high-pass kernel on a data set containing noise. The elevation data set shown in the top portion of the figure contains a number of anomalous cells. These cells have extremely high values (spikes, shown in black) or low values (pits, shown in white) relative to nearby cells. If uncorrected, pits and spikes will affect slope, aspect, and other terrain-based calculations. These locally extreme values should be identified and modified.

The high-pass kernel shown contains a value of 9 in the center and -1 in all other cells. Each value is divided by 9 to reduce the range of the output variable. The kernel returns a value near the local average in smoothly changing areas. The positive and negative values balance, returning small numbers in flat areas.

The high-pass kernel generates a large positive value when centered on a spike. The large differences between the center cell and adjacent cells are accentuated. Conversely, a large negative value is generated when a pit is encountered. An example shows the application of the high-pass filter for a cell near the upper left corner of the input data layer (Figure 10-21). Each cell value is multiplied by the corresponding kernel coefficient. These numbers are summed, and divided by 9, and the result placed in the corresponding output location. Calculation results are shown as real numbers, but cell values are shown here recorded as integers. Output values may be real numbers or integers, depending on the programming algorithm and perhaps the specifications set by the user.

The mean filter is representative of many moving window functions in that it increases the *spatial covariance* in the output data set. High spatial covariance means values are autocorrelated (discussed in greater depth when in the spatial prediction section of Chapter 13). A large positive spatial covariance means cells near each other are likely to have similar values. Where you find one cell with a large number, you are likely to find more cells with large numbers. If spatial data have high spatial covariance, then small numbers are also likely to be found near each other. Low spatial covariance means nearby values are unrelated – knowing the value at one cell does not pro-

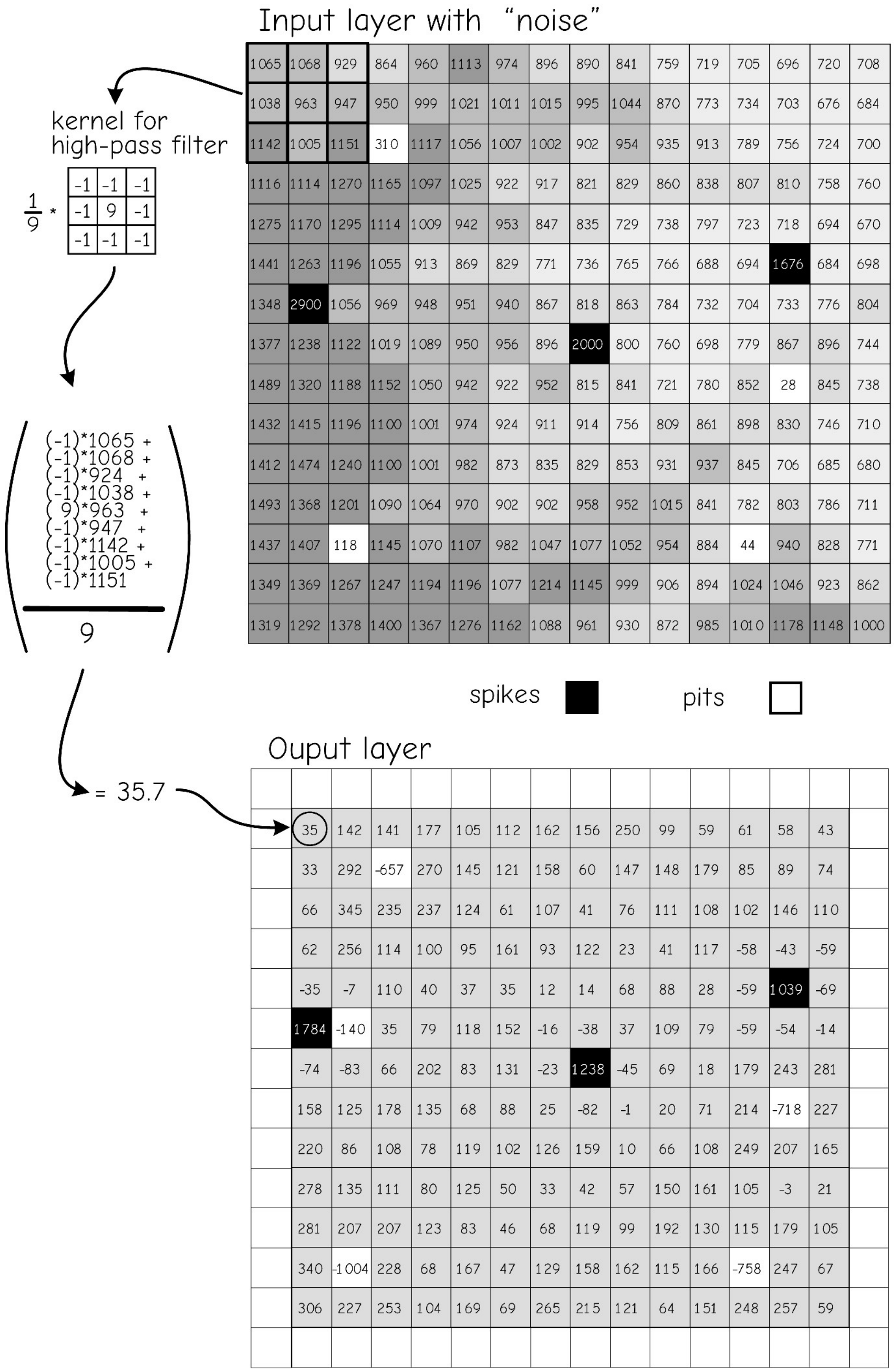


Figure 10-21: An example of a moving window function. Raster data often contain anomalous "noise" (dark and light cells). A high-pass filter and kernel, shown at top left, highlights "noisy" cells. Local differences are amplified so that anomalous cells are easily identified.

vide much information about the values at nearby cells. High spatial covariance in the "real world" may be a good thing. If we are prospecting for minerals, then a sample with a high value indicates we are probably near a larger area of ore-bearing deposits. However, if the spatial autocorrelation is increased by the moving window function, we may get an overly optimistic impression of our likelihood of striking it rich.

The spatial covariance increases with many moving window functions because these functions share cells in adjacent calculations. Note the average function in Figure 10-22. The left of Figure 10-22 shows sequential positions of a 3 by 3 window. In the first window location, the mean is calculated and placed in the output layer. The window center is then shifted one cell to the right, and the mean for this location calculated and placed in the next output cell to the right. Note that there are six cells in common for these two means. Adjacent cells in the output data layer share six of nine cells in the mean calculation. When a particularly low or high cell occurs, it affects the mean of many cells in the output data layer. This causes the outputs to be quite similar, and increases the spatial covariance.

Zonal Functions

Zonal functions apply operations based on defined regions, or zones, within an area. Typically, the zones are recorded in a data layer, with a unique identifier for each zone. A function is then applied based on the zone.

There are many reasons for applying zonal functions. We often want to summarize data for defined units in a region, including total population in a county, average rainfall in a watershed, or number of impoverished families across neighborhoods. More complicated analyses may require different operations be applied to different zones; for example, we may be creat-

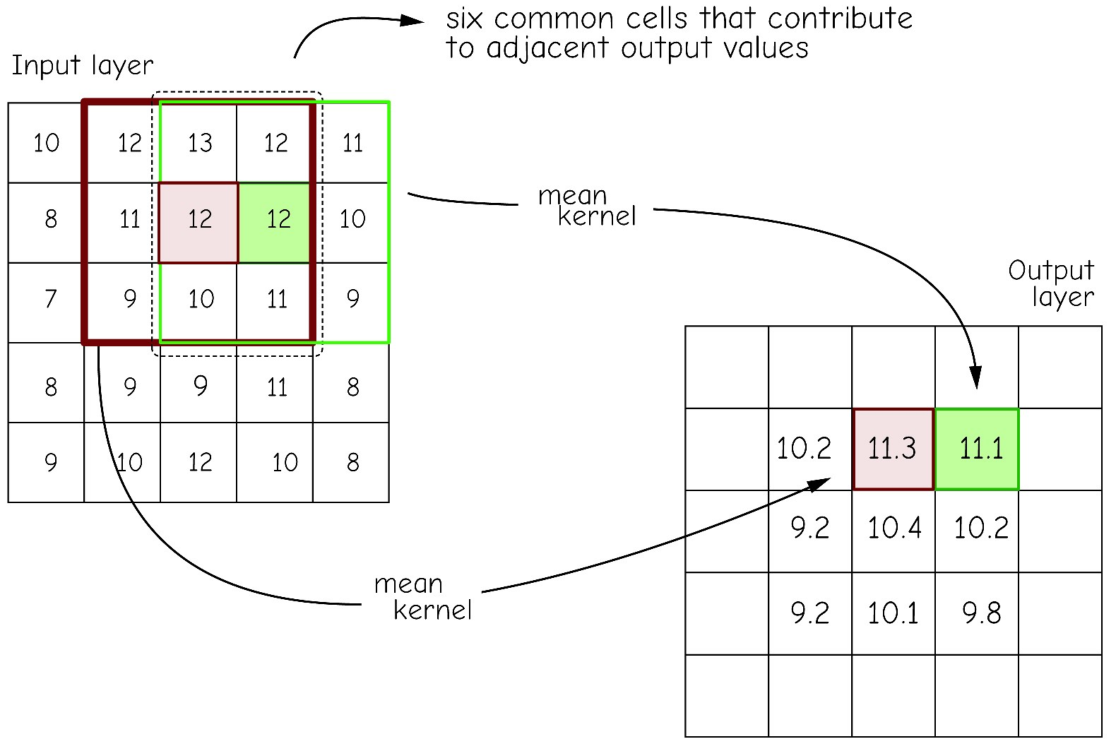


Figure 10-22: A moving window may increase spatial covariance. Adjacent output cells share many input cell values. In the mean function shown here, this results in similar output cell values.

ing an elevation data set from many sources, and we may wish to use the highest-quality data in zones where it exists, and use successively poorer data in other zones. Zonal functions give us these capabilities.

Figure 10-23 illustrates the application of a zonal function. In this example, the function calculates the zonal average for In_Layer, based on zones defined by Zone_Layer. The syntax here is

```
Out_Layer =
    ZoneAvg(In_Layer, Zone_Layer)          (10.8)
```

There is no standard syntax across software, so the specific order and interpretation of operands depend on the software used.

Note that the output here is a raster, with identical values in all the cells of a given zone. This is typically how zonal functions are specified. Most systems can create a table with zonal identifiers and summary values to accompany the layer, and in some cases, the operation only results in a table.

Zonal functions typically require compatible cell sizes. Generally, this means the cells in the input layers and zone-defining layer have the same cell size and orientation. The zone layer may have dimensions that are integer multiples of the input layers, but the reverse is generally not recommended. Input layer cell sizes larger than zone layer cell sizes may lead to ambiguous zone definition when more than one zone may correspond to an input cell.

Cost Surfaces

Many problems require an analysis of travel costs. These may be monetary costs of travel, such as the price one must charge to profitably deliver a package from the nearest distribution center to all points in a region. Travel costs might also be measured in other units, for instance, the time it takes to travel from a school to the nearest hospital, or as a likelihood, such as the chance of a noxious foreign weed spreading out from an introduction point. These analyses may be performed with the help of *cost surfaces*.

Out_Layer = ZoneAvg (In_Layer, Zone_Layer)

In_Layer

7	4	2	3	6	10
1	3	1	4	1	9
1	4	11	10	8	7
3	1	2	7	8	10
2	2	11	7	10	12
5	8	10	9	13	11

Zone_Layer

a	a	a	b	b	b
a	a	a	b	b	b
a	a	c	b	c	b
a	b	b	b	c	c
a	a	b	c	c	c
b	b	c	c	c	c

Out_Layer

2.7	2.7	2.7	6.6	6.6	6.6
2.7	2.7	2.7	6.6	6.6	6.6
2.7	2.7	9.2	6.6	9.2	6.6
2.7	6.6	6.6	6.6	9.2	9.2
2.7	2.7	6.6	9.2	9.2	9.2
6.6	6.6	9.2	9.2	9.2	9.2

Figure 10-23: An example of a zonal function. Averages are calculated based on the zones stored in Zone_Layer.

A cost surface contains the minimum cost of reaching cells in a layer from one or more source cells (Figure 10-24). The simplest cost surface is based on a uniform travel cost. Travel cost depends only on the distance covered, with a fixed cost applied per unit distance traveled. This cost per unit distance does not change from cell to cell. There are no barriers, so the straight line distance is converted to a cost. First, the distance is calculated from our source or starting location to each cell. As illustrated in Figure 10-24, the distance is calculated based on the Pythagorean formula. Distances to each cell in the x and y directions contribute to the total distance from a source cell or cells.

The distance from a source cell is combined with a fixed cost per unit distance to calculate travel cost. As shown in the right side of Figure 10-24, each distance value is multiplied by the fixed cost factor. This results in a cost surface, a raster layer containing the travel cost to each cell. If there are multiple source cells, travel costs are calculated from each source cell, and the lowest cost is typically placed in the output cell.

Note that distance is commonly measured at least two ways — a straight line (Euclidian) distance, as shown in Figure 10-24, or as a row-column distance. A row-column distance is measured along the row and column axes, and is by definition longer than the straight-line distance. Straight line distances are preferred in most applications, although they are more difficult to implement.

Travel costs may also be calculated using a *friction surface*. The cell values of a friction surface represent the cost per unit travel distance for crossing each cell. Friction surfaces are used to represent areas with a variable travel cost. Imagine a large military base. Part of the base may include flat,

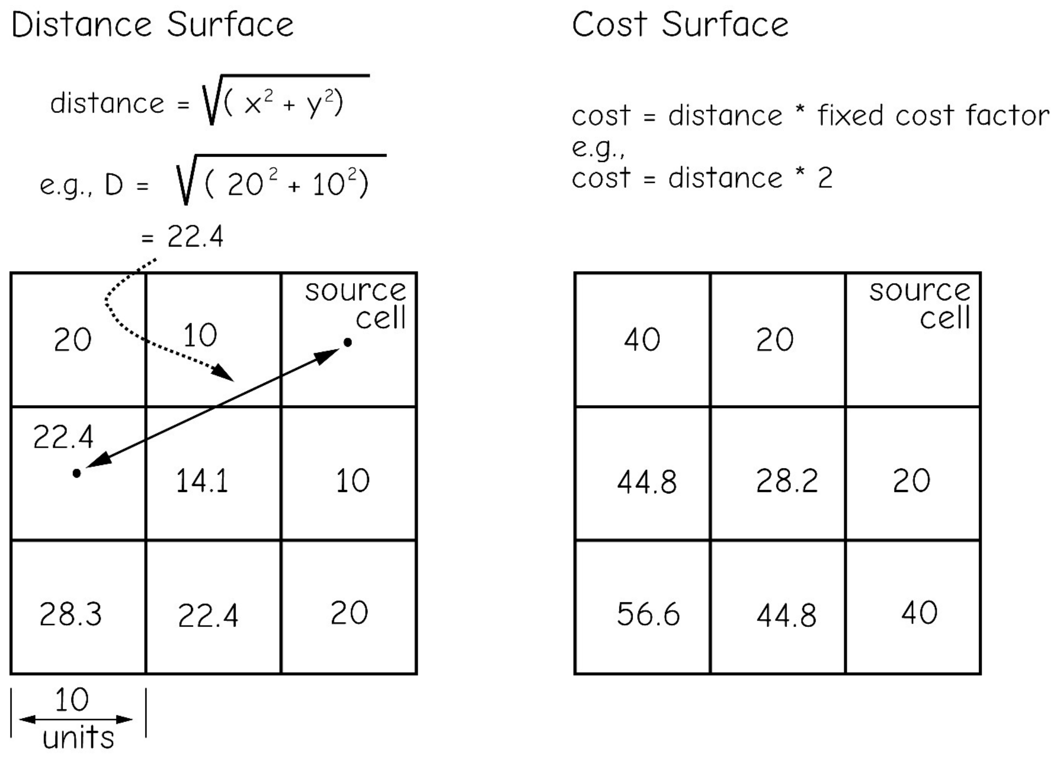


Figure 10-24: A cost surface based on a fixed cost per unit distance. Minimum distance from a set of source cells is multiplied by a fixed cost factor to yield a cost surface.

smooth areas such as drill fields, parking lots, or parade grounds. These areas are relatively easy to cross, with correspondingly low travel times per unit distance. Other parts of the base may be covered by open grasslands. While the surface may be a bit rougher, travel times are still moderate. Other parts may be composed of forests. These areas would have correspondingly high travel times, as a vehicle would have to pick a path among the trees. Finally, there may be areas occupied by water, fences, or buildings. These areas would have effectively infinite travel times.

Each cell in the friction surface contains the cost required to traverse a portion of the cell (Figure 10-25). A value of 3 indicates it costs three units (of time, money, or other factor) per unit distance in the cell. If a cell is 10 wide and costs 3 units per unit distance, and the cell is crossed along the width, then the cost for traversing the cell is 10 times 3, or 30 units.

The actual cost for traversing the cell depends on the distance traveled through the cell. When a cell is traversed parallel to the row or column edge, then the distance is simply the cell dimension. When a cell is traversed at any other angle, the distance will vary. It may be greater or less than the cell dimension, depending on the angle and location of the path.

The travel cost required to reach each cell is the minimum accumulated total of the cost times the distance to a source cell. We specify a minimum accumulated cost because if there is more than one source cell, there is a large number of potential paths to each of these source cells. Distance across each cell is multiplied by the friction surface cost for that cell and summed for a path to accumulate the total travel cost. The lowest cost path from a source location to a cell is usually assigned as the travel cost to that cell.

Figure 10-25 shows an example of calculations for the friction cost along a set of paths. These are straight line paths that travel either parallel to the cell boundaries (purely in an x or y direction) or at some angle across cells.

Sample calculation of the friction costs for a path parallel to the x axis is shown at the top middle and on the left side of Figure 10-25. Note that when traveling parallel to a cell boundary, one half-cell width is traversed in the starting and ending cells. Intermediate cells are crossed at a full cell width. When moving from the starting cell to the adjacent left cell, a friction surface value of

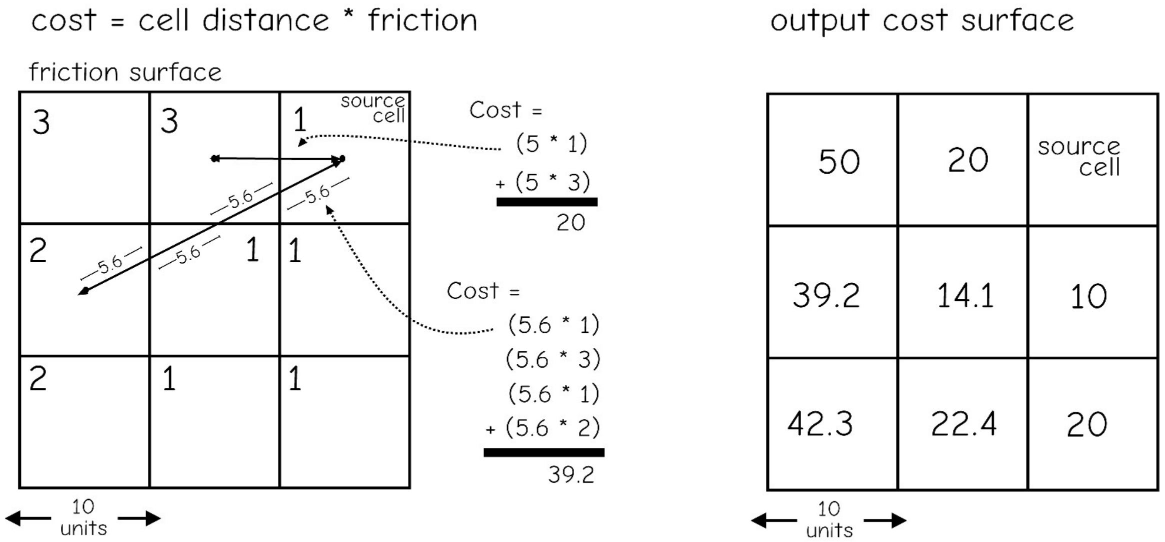


Figure 10-25: A cost surface based on spatially variable travel costs. A friction surface specifies the spatially varying cost of traveling through raster cells. The distance traversed through each cell is multiplied by the cost in the friction surface. The values are summed for each path to yield a total cost.

1 is encountered, then a friction surface value of 3. One-half the distance, 5 units, is through the top-right cell at a per-unit friction cost of 1. One-half the distance is through the adjacent cell to the left, at a per-unit friction cost of 3. The total cost is then the distance traveled in each cell multiplied by the per-unit friction cost of the cell:

$$5 * 1 + 5 * 3 = 20 \quad (10.9)$$

The friction cost when traversing cells at an angle is illustrated at the bottom left and bottom center of Figure 10-25. The friction cost is the sum of the cell cost per unit distance multiplied by the distance traveled in each cell. The path begins at the source cell and ends two cells to the left and one cell down. Each intervening cell is traversed for a distance of 5.6 cell units. The distance traversed in each cell is multiplied by the friction value for each cell. The total cost for this leg is:

$$5.6*1 + 5.6*3 + 5.6*1 + 5.6*2 = 39.2 \quad (10.10)$$

In general, the cost of any path is expressed as:

$$Totalcost = d_1*c_1+d_2*c_2+....d_n*c_n \quad (10.11)$$

where d_i is the distance and c_i is the cost across each cell of a path.

Many softwares calculate the cumulative cost for the most direct path using a slightly different approach, called the *row-column distance*. Rather than travel along a straight line path, the row-column distance travels from cell center to cell center (Figure 10-26). Calculations are much easier because the length of the path within each cell is constant with row column distance, and for square cells this distance equals the cell width (or height). The distance in each cell varies when using the straight line distance, and so the time required to calculate the accumulated distance is substantially increased. The row/column distance gives the same relative costs for travel from a source cell to each target cell, but the absolute costs change (compare the costs on the right of Figure 10-25 to the right of Figure 10-26).

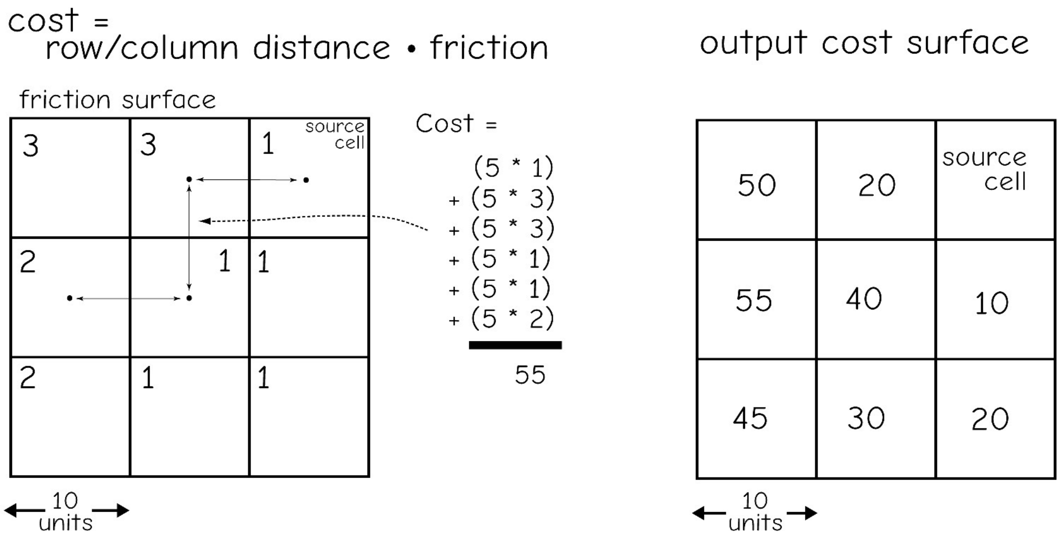


Figure 10-26: Calculations of the travel cost from a source cell to each other cell using row-column distance calculations.

Many implementations of a friction surface or cost function allow you to search for the minimum cost to travel to a cell from a set of source cells. The straight line distance may not be the "least costly," and so alternatives may be examined. There are many routes from any source cell to any destination cell, thousands of distinct routes in most instances. Software typically implements some optimization algorithm to eliminate routes early on and reduce search time, thereby arriving at the cost surface in some acceptable time period.

Note that barriers may be placed on a cost surface to preclude travel across portions of the surface. These barriers may be specified by setting the cost so high that no path will include them. Any circuitous route will be less expensive than traveling over the barriers. Some software allows the specification of a unique code to identify barriers, and this code precludes movement across the cell.

Summary

Raster analyses are essential tools in GIS, and should be understood by all users. Raster analyses are widespread and well developed for many reasons, in part due to the simplicity of the data structure, the ease with which continuous variables may be represented, and the long history of raster analyses.

Map algebra is a concept in which raster data layers are combined via summation and multiplication. Values are combined on a cell-by-cell basis, and may be added, subtracted, multiplied, or divided. Care must be taken to avoid ambiguous combinations in the output that originate from distinct input combinations.

Raster analyses can be local, neighborhood, or global, and general analyses such as buffering and overlay may be applied using raster data sets. Neighborhood operations are particularly common in raster analyses, and may be applied with a moving window approach. A moving window is swept across all cells in a data layer, typically multiplying kernel values by data found around a center cell. Window size and shape may be modified at the edges of the data layers. Moving windows may be used to specify a wide range of combinatorial, terrain, and statistical functions.

Cost or friction surfaces are an important subset of proximity analyses that may be easily applied in raster analyses. A cost surface identifies the travel costs required for movement from a specified set of locations.

Suggested Reading

Berry, J.K. (1986). A mathematical structure for analyzing maps. *Environmental Management*, 11:317–325.

Berry, J.K. (1987). Fundamental operations in computer-assisted mapping. *International Journal of Geographic Information Systems*, 1:119–136.

Bonham-Carter, G.F. (1996). *Geographic Information Systems for Geoscientists: Modelling with GIS*. Ottawa: Pergamon.

Burrough, P.A., McDonnell, R.A. (1998). *Principles of Geographical Information Systems* (2nd ed.). New York: Oxford University Press.

Cliff, A.D., Ord, J.K. (1987). *Spatial Autocorrelation*. New York: Methuen.

DeMers, M.N. (2002). *GIS Modeling in Raster*. New York: Wiley.

de Smith, M.J., Goodchild, M.F., Longley, P.A. (2007). *Geospatial Analysis, a Comprehensive Guide to Principles, Techniques, and Software Tools*. Leicester: Matador.

Eastman, J.R., Jin, W., Keym, P.A.K., Toledano, J. (1995). Raster procedures for multi-criteria/multi-objective decisions. *Photogrammetric Engineering and Remote Sensing*, 61:539–547.

Eastman, J.R. (1997). *Idrisi for Windows*. Worcester: Clark University.

Hengl, T. (2006). Finding the right pixel size. *Computers and Geosciences*, 32:1283–1298.

Mitchell, A. (1999). *The ESRI Guide to GIS Analysis: Geographic Patterns and Relationships*. Redlands: ESRI Press.

Morain, S., Baros, S.L. (1996). *Raster Imagery in Geographic Information Systems*. Santa Fe: OnWord Press.

Tomlin, C. D. (1990). *Geographic Information Systems and Cartographic Modeling*. Upper Saddle River: Prentice-Hall.

Study Questions

10.1 - What is map algebra?

10.2 - Why must raster layers have compatible cell sizes and orientations for most raster combination operations?

10.3 - What is a null value in a raster data set? How is this null value typically treated in a raster operation?

10.4 - Perform the listed raster operations.

3	2	4	11	9	1	3
1	6	5	20	14	8	7
7	13	2	1	4	9	11
12	11	10	8	5	6	10
3	2	1	17	12	11	9
8	5	6	8	3	13	16
19	17	9	11	12	7	15

Perform the following operations with a 3x3 window, centered on the noted cells:

-average, on the pentagon,
-standard deviation, on the circle,
-maximum, on the triangle,
-value range, on the square,
-average, on the ellipse,
-median, on the star

10.5 - Perform the listed raster operations.

3	2	4	11	9	1	3
1	6	5	20	14	8	7
7	13	2	1	4	9	11
12	11	10	8	5	6	10
3	2	1	17	12	11	9
8	5	6	8	3	13	16
19	17	9	11	12	7	15

Perform the following operations with a 3x3 window, centered on the noted cells:

-average, on the pentagon,
-standard deviation, on the circle,
-maximum, on the triangle,
-value range, on the square,
-average, on the ellipse,
-median, on the star

10.6 - What are the values in cells C1, C2, C3, and C12 in the output layer?

Con(Layer1 < 2, 0, 1)

Layer1

2	N	1	2
1	N	1	3
4	1	2	0
N	2	N	1

Output

C1	C2	C3	C4
C5	C6	C7	C8
C9	C10	C11	C12
C13	C14	C15	C16

10.7 - What are the values in cells C5, C7, C10, and C13 in the output layer?

Con(Layer1 < 2, 0, 1)

Layer1

2	N	1	2
1	N	1	3
4	1	2	0
N	2	N	1

Output

C1	C2	C3	C4
C5	C6	C7	C8
C9	C10	C11	C12
C13	C14	C15	C16

10.8 - What are the cell values for cells C1, C3, C4, and C10 in the output layer, below?

Output = CON((layerA==N), 1, layerA)

layerA

N	N	1	0
1	N	2	N
N	4	N	N
0	1	N	1

Output

C1	C2	C3	C4
C5	C6	C7	C8
C9	C10	C11	C12
C13	C14	C15	C16

10.9 - What are the cell values for cells C2, C5, C7, and C11 in the output layer, below?

Output = CON((layerA==N), 1, layerA)

layerA

N	N	1	0
1	N	2	N
N	4	N	N
0	1	N	1

Output

C1	C2	C3	C4
C5	C6	C7	C8
C9	C10	C11	C12
C13	C14	C15	C16

10.10 - Give an example of a nested operation.

10.11 - What are the values in output cells C9, C10, C11, and C12?

Output = CON(ISNULL(layerA), 1, N)

layerA

1	N	N	0
0	N	2	1
N	1	5	0
N	1	N	1

Output

C1	C2	C3	C4
C5	C6	C7	C8
C9	C10	C11	C12
C13	C14	C15	C16

10.12 - What are the values in output cells C7, C8, C13, and C16?

Output = CON(ISNULL(layerA), 1, N)

layerA

1	N	N	0
0	N	2	1
N	1	5	0
N	1	N	1

Output

C1	C2	C3	C4
C5	C6	C7	C8
C9	C10	C11	C12
C13	C14	C15	C16

10.13 - What is the scope of a raster operation?

10.14 - Does a NOT operation applied to a raster cell value containing a NULL value return a NULL value, a zero value, a 1, or some other non-null value?

10.15 - Diagram an AND operation on a raster data cell.

10.16 - Provide the answer for the following logical operations:

1	1	0	0
0	0	0	1
1	1	0	0
1	0	0	1

and

0	1	0	1
1	0	0	1
1	1	1	0
1	0	1	1

=

1	1	0	0
0	0	0	1
1	1	0	0
1	0	0	1

or

0	1	0	1
1	0	0	1
1	1	1	0
1	0	1	1

=

10.17 - Provide the answer for the following logical operations:

0	0	N	3
1	0	0	3
1	7	1	0
0	N	0	1

and

0	5	0	1
1	0	0	1
1	0	1	0
1	0	1	1

=

0	1	0	0
0	0	0	1
3	0	0	0
1	0	1	1

or

0	1	0	9
1	0	0	1
1	1	N	0
0	0	6	0

=

10.18 - Describe how local arithmetic functions can be used to apply a clip function in a raster environment.

10.19 - What is a kernel in a moving window operation? Does the kernel size or shape change for different portions of the raster data set? Why or why not?

10.20 - What moving window operation would most likely use the kernel below?

-2	-1	-2	-1	-2
-1	0	0	0	-1
-2	0	25	0	-2
-1	0	0	0	-1
-2	-1	-2	-1	-2

10.21 - What moving window operation would most likely use the kernel below?

1	2	1
0	0	0
-1	-2	-1

10.22 - What is meant by high spatial covariance in a raster data layer?

10.23 - Calculate the cost of travel between A and B, and A and C, over the cost surface below, both by straight line, and by row-column paths.

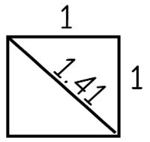
Source/target cells

A			
B			C

↕10 units

Cost surface

3	5	6	8
4	1	7	5
2	5	1	6
2	4	1	1



remember, the diagonal of a square is 1.41 x the edge

10.24 - Calculate the cost of travel between A and B, and A and C, over the cost surface below, both by straight line, and by row-column paths:

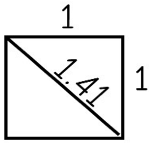
Source/target cells

A			
B			C

↕10 units

Cost surface

2	4	7	8
3	1	7	9
5	1	4	7
1	4	1	2



remember, the diagonal of a square is 1.41 x the edge

11 Terrain Analysis

Introduction

Elevation and related terrain variables are important at some point in almost everyone's life. Elevation and slope change across the landscape (Figure 11-1), and this variation determines where rivers flow, lakes occur, and floods are frequent. Terrain variation influences soil moisture and hence food production. Terrain in large part affects water quality through sediment generation and transport. Terrain strongly influences transportation networks and the cost and methods of building construction. Terrain

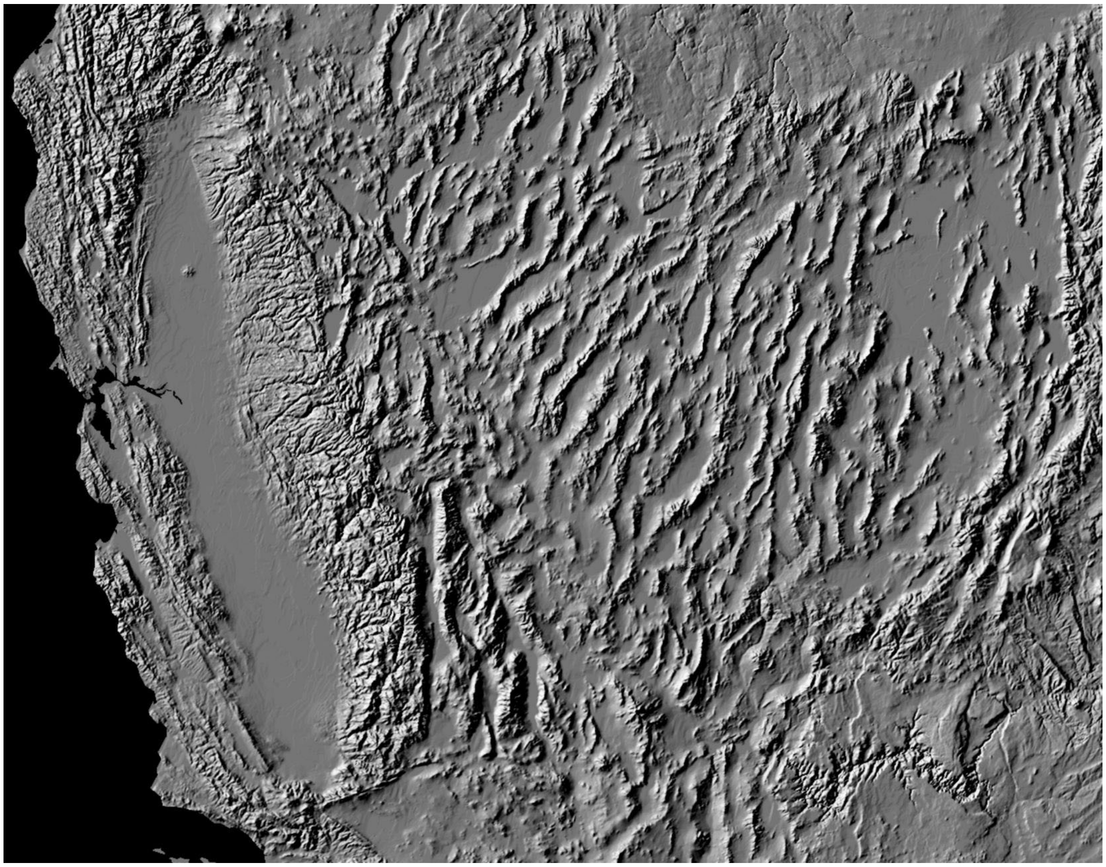


Figure 11-1: An example of a terrain-based image of the western United States. Shading, based on local elevation, emphasizes terrain shape. Topographic features are clearly identified, including the Central Valley of California at the left of the image, and to the center and right, the parallel mountains and valleys of the Basin and Range region (courtesy USGS).

variables are frequently applied in a broad range of spatial analyses (Table 11-1).

Given the importance of elevation and other terrain variables in resource management, and the difficulties of manual terrain analysis, it is not surprising that terrain analysis is well developed in GIS. Indeed, it is often impractical to perform consistent terrain analyses without a GIS. For example, slope calculations over large areas based on manual methods are slow, error prone, and inconsistent. Elevation change over a horizontal distance is difficult to measure, these measurements are slow, and estimates are likely to vary among human analysts. In contrast, digital slope calculations are easy to program, consistent, and have proven to be as accurate as field measurements.

Both data and methods exist to extract important terrain variables via a GIS. Digital elevation models (DEMs), described in Chapters 2 and 7, have been developed for most of the world using methods described in Chapters 5 and 6, and DEM renewal and improvement continues.

Table 11-1: A subset of commonly used terrain variables (adapted from Moore et al., 1993).

Variable	Description	Importance
Height	Elevation above base	Temperature, vegetation, visibility
Slope	Rise relative to horizontal distance	Water flow, flooding, erosion, travel cost, construction suitability, geology, insolation, soil depth
Aspect	Downhill direction of steepest slope	Insolation, temperature, vegetation, soil characteristics and moisture, visibility
Upslope area	Watershed area above a point	Soil moisture, water runoff volume and timing, pollution or erosion hazards
Flow length	Longest upstream flow path to a point	Sediment and erosion rates
Upslope length	Mean or total upstream flow path length from a point	Sediment and erosion rates
Profile curvature	Curvature parallel to slope direction	Erosion, water flow acceleration
Plan curvature	Curvature perpendicular to slope direction	Water flow convergence, soil water, erosion
Visibility	Site obstruction from given viewpoints	Utility location, viewshed preservation

Calculations are based on cell values assigned to a regular grid. We use the concept of Z values, the height stored in the raster arrays, to extract information about terrain, using the magnitudes and patterns of changes in Z across the grid (Figure 11-2). For example, the height differences between adjacent cells or in a neighborhood of cells are used to calculate a local slope (slope in Figure 11-2). The angle and orientation of lines defined by x, y, and Z values near a point are used to calculate the normal vector, at right angles to the local surface (Figure 11-2). Local curvature and slope direction are also calculated by differences in Z values in a neighborhood.

Many terrain analysis functions can be specified by a mathematical operation applied to an appropriate moving window. The results from these mathematical operations in turn provide important information about terrain characteristics that are helpful in spatial analysis.

Slope and Aspect

Slope and aspect are two commonly used terrain variables. They are required in many studies of hydrology, conservation, site planning, and infrastructure development, and are the basis for many other terrain analysis functions. Road construction costs and safety are sensitive to slope. Watershed boundaries, flowpaths and direction, erosion modeling, and viewshed determination (discussed later in this chapter) all use slope and/or aspect data as input. Slope or aspect may be useful in mapping both vegetation and soil resources.

Slope is defined as the change in elevation (a rise) with a change in horizontal position (a run). Seen in cross section, the slope is related to the rise in elevation over the run in horizontal position (Figure 11-3). Slope is often reported in degrees, between zero (flat), and 90 (vertical). The slope is equal to 45 degrees when the rise is equal to the run. The slope in degrees is calculated from the rise and run through the tangent trigonometric function. By definition, the tangent of the

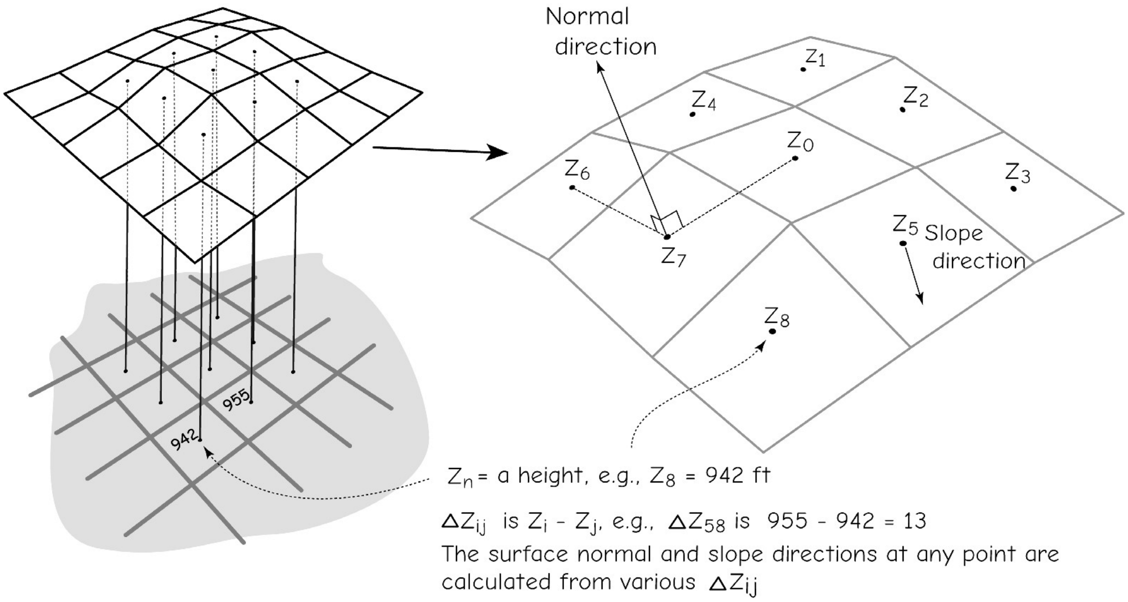


Figure 11-2: A depiction of a surface represented by a raster DEM, and changes in Z values for cells used for calculating various terrain attributes.

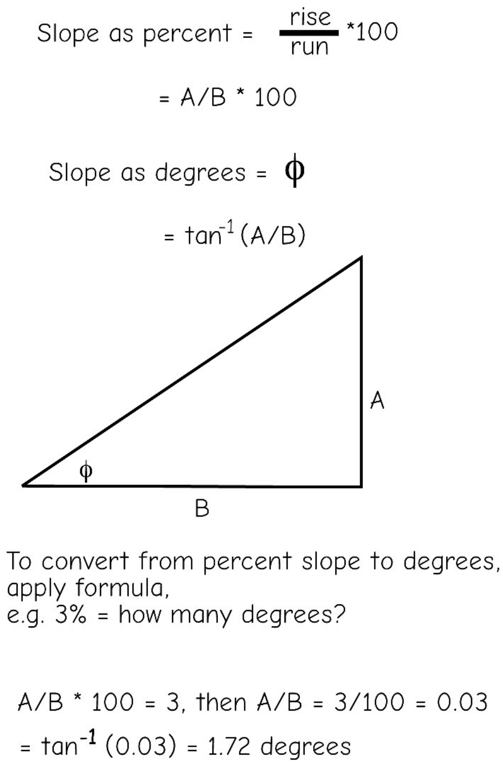


Figure 11-3: Slope formula, showing the rise (A), run (B), and slope angle (ϕ).

slope angle (ϕ) is the ratio of the rise over the run, as shown in (Figure 11-3). The inverse tangent of a measured rise over a run gives the slope angle. A steeper rise or shorter run lead to a higher ϕ and hence steeper slope.

Slope may also be expressed as a percent, calculated by 100 times the rise over the run (Figure 11-3). Slopes expressed as a percent have magnitudes from zero (flat) to infinite (vertical), with a sign convention inconsistently defined. Some authors define a positive slope as uphill (0 to +∞), and a negative slope downhill (0 to -∞), while others define slope only downhill (0 to +∞). A slope of 100% occurs when the rise equals the run.

Calculating slope from a raster data layer is more complicated than in the cross-section view shown in Figure 11-3. The raster cells occur at regular intervals across an irregular terrain surface. Slope direction at a point in the landscape is typically measured in the steepest direction of elevation change

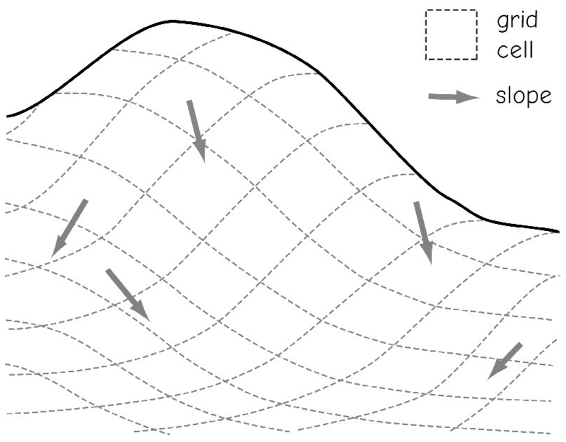


Figure 11-4: Slope direction, shown as gray arrows for some example locations above, often changes substantially among cells on a raster surface. Slope calculations in three dimensions require the consideration of all values surrounding a cell.

(Figure 11-4). Slope changes in a complex way across many landscapes, and calculations of slope must factor in the relative changes in elevations around a central cell.

As demonstrated in Figure 11-4, the slope direction often does not point parallel to the raster rows or columns. Consider the cells depicted in Figure 11-5. Higher eleva-

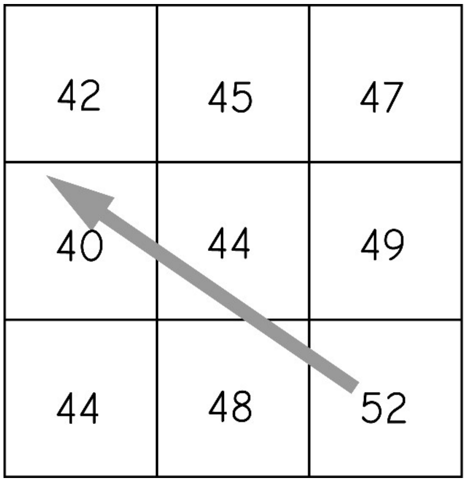


Figure 11-5: Slope direction on a raster surface usually does not point from cell center to cell center. Therefore, formulae that accurately represent slope on a surface integrate several cells surrounding the center cell.

tions occur at the lower right corner, and lower elevations occur toward the upper left. The direction of steepest slope trends from one corner towards the other, but does not pass directly through the center of any cell. How do we obtain values for the rise and run? Which elevations should be used to calculate slope? Intuitively we should use some combination of a number of cells in the vicinity of the center cell, perhaps all of them.

Elevation is often represented by the letter Z in terrain functions. These terrain functions are usually calculated with a symmetrical moving window. A 3 by 3 cell window is most common, although 5 by 5 and other odd-numbered windows are also used. Each cell in the window is assigned a subscript, and the elevation values found at window locations referenced by subscripted Z values.

Figure 11-6 shows an example of a 3 by 3 cell window. The central cell has a value of 44, and is referred to as cell Z_0. The upper left cell is referred to as Z_1, the upper center cell as Z_2, and so on through cell Z_8 in the lower right corner.

Slope at each center cell is most commonly calculated from the formula:

$$s = \operatorname{atan}\sqrt{\left(\frac{dZ}{dx}\right)^2 + \left(\frac{dZ}{dy}\right)^2} \qquad (11.1)$$

where s is slope, atan is the inverse tangent function, Z is elevation, x and y are the respective coordinate axes, and dZ/dx and dZ/dy are calculated for each cell based on elevation values surrounding a given cell. The symbol dZ/dx represents the rise (change in Z) over the run in the x direction, and dZ/dy represents the rise over the run in the y direction. These formulas are combined to calculate the slope for each cell based on the combined change in elevation in the x and y directions.

Many different formulas and methods have been proposed for calculating dZ/dx and dZ/dy. The simplest, shown in Figure 11-6 and at the top of Figure 11-7, use the four cells nearest to Z_0.

$$dZ/dx = (Z_5 - Z_4)/(2C) \qquad (11.2)$$

$$dZ/dy = (Z_2 - Z_7)/(2C) \qquad (11.3)$$

where C is the cell dimension and the Zs are defined as in Figure 11-6. This method uses the "four nearest" cells, Z_4, Z_5, Z_2, and Z_7, in calculating dZ/dx and dZ/dy. These four cells share the largest common border with the center. This four nearest method is perhaps the most obvious and provides reasonable slope values under many circumstances.

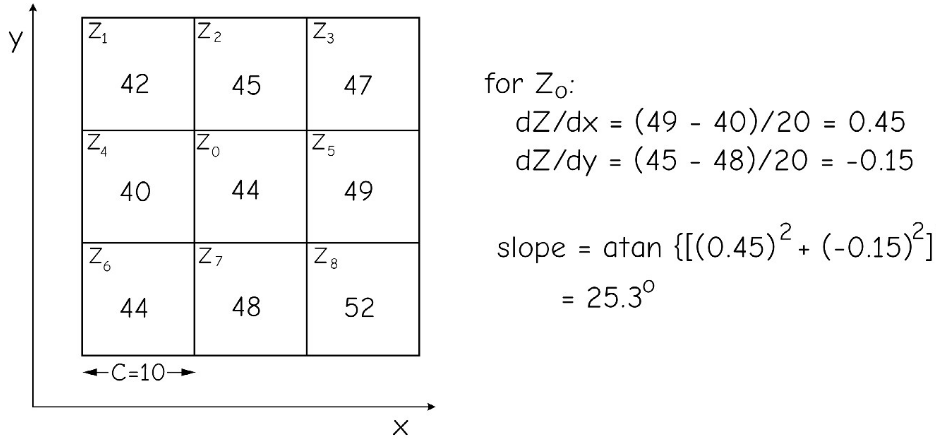


Figure 11-6: Slope calculation based on cells adjacent to the center cell.

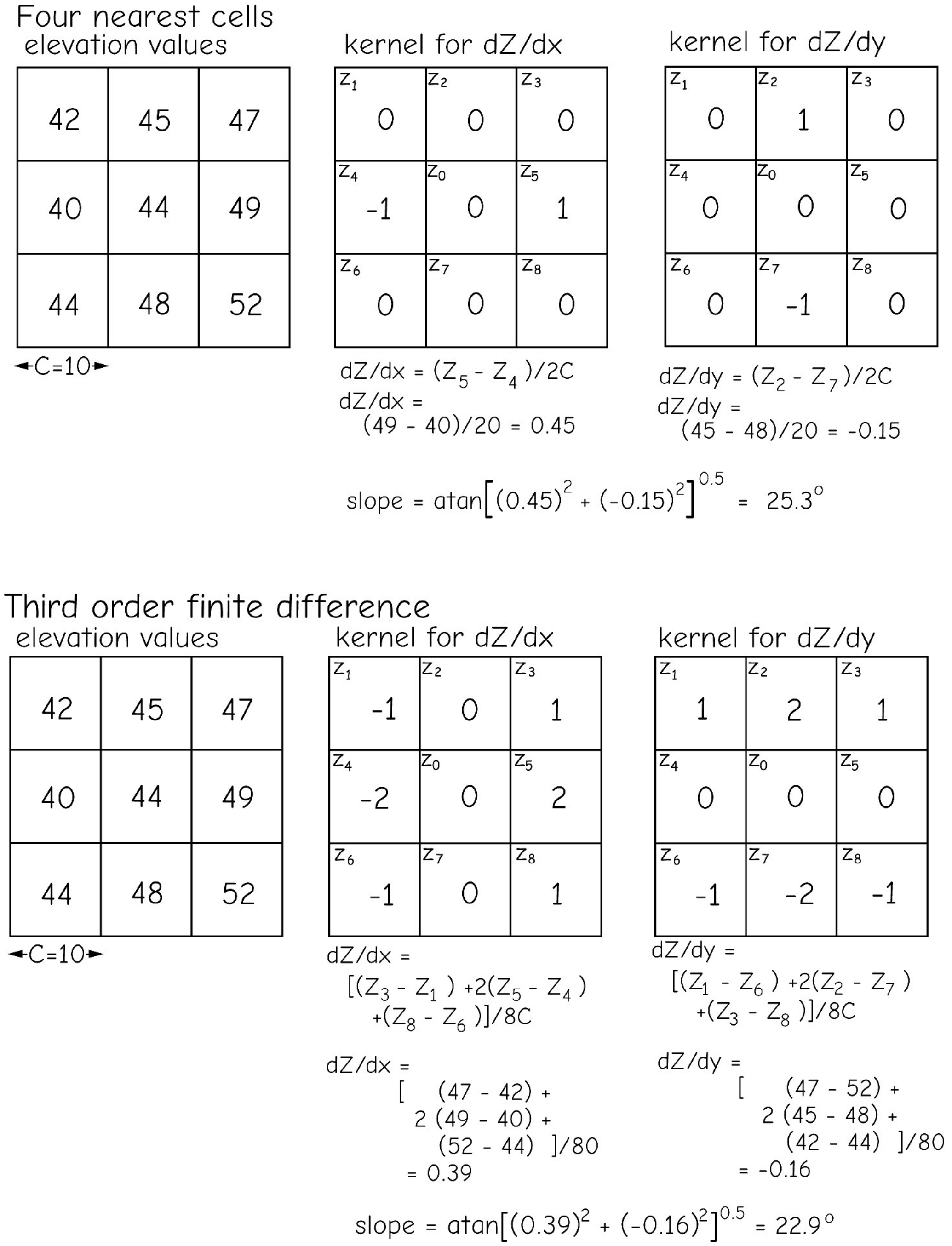


Figure 11-7: Four nearest cells method (top) and third order finite difference method (bottom, explained on the next page), used in calculating slope. C is cell size and dZ/dx and dZ/dy are the changes in elevation (rise) with changes in horizontal position (run). Note that different slope values are produced by the different methods.

A common alternate method is known as a *third order finite difference* approach (Figure 11-7, bottom). This method for calculating dZ/dx and dZ/dy differs mainly in the number and weighting it gives to cells in the vicinity of the center cell. The four nearest cells are given a higher weight than the "corner" cells, but data from all eight nearest cells are used.

Several other methods have been developed that are better for calculating slope under certain conditions. Better means that, on average, a method produces more accurate slope estimates when compared to carefully collected field measurements. However, no method has proved best under all terrain conditions. Literature on the methods, their derivation, and application are listed at the end of this chapter.

Comparative studies have shown the two methods described here to be among the best for calculating slope and aspect over a wide range of conditions. The method using the four nearest cells was among the best for smooth terrain, and the 3rd order finite difference approach is often among the best when applied to rough terrain.

Aspect is also an important terrain variable that is commonly derived from digital elevation data. The aspect at a point is the steepest downhill direction. The direction is typically reported as an azimuth angle, with zero in the direction of grid north, and the azimuth angle increasing in a clockwise direction (Figure 11-8). Aspects defined this way take values between 0 and 360 degrees. Flat areas have no aspect, because there is no downhill direction.

Aspect (α) is most often calculated using dZ/dx and dZ/dy:

$$\alpha = 180 - \operatorname{atan}\left(\frac{\left(\frac{dZ}{dy}\right)}{\left(\frac{dZ}{dx}\right)}\right) + 90\left(\frac{\left(\frac{dZ}{dx}\right)}{\left|\frac{dZ}{dx}\right|}\right) \qquad (11.4)$$

where atan is the inverse tangent function that returns degrees, and dZ/dy and dZ/dx are defined as above.

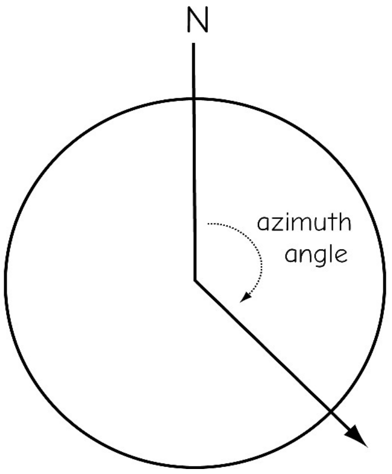
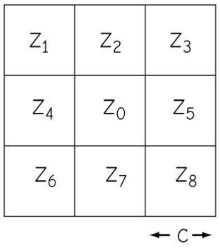


Figure 11-8: Aspect may be reported as an azimuth angle, measured clockwise in degrees from north.

As with slope calculations, estimated aspect varies with the methods used to determine dZ/dx and dZ/dy. Tests have shown the four nearest cell and third order finite difference methods again yield among the most accurate results, with the third order method among the best under a wide range of terrain conditions.

Profile curvature and *plan curvature* are two other local topographic indices that are important in terrain analysis and may be derived from gridded elevation data. Profile and plan curvature are helpful in measuring and predicting soil water content, overland flow, rainfall-runoff response in small catchments, and the distribution of vegetation.

Profile curvature is an index of the surface shape in the steepest downhill direction (Figure 11-9). The profile curvature may be envisioned by imagining a vertical plane, slicing downward into the earth surface, with the plane containing the line of steepest descent (aspect direction). The surface traces a path along the face of this plane, and the curvature is defined by the shape of this path. Smaller values of profile curvature indicate a concave (bowl shaped) path in the downhill direction, and larger values of profile curvature indicate a convex (peaked) shape in the downhill direction.



$$D = [\,(Z_4 + Z_5)/2 - Z_0\,] \;/\; C^2$$

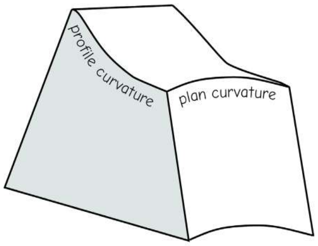
$$E = [\,(Z_2 + Z_7)/2 - Z_0\,] \;/\; C^2$$

$$F = (Z_3 - Z_1 + Z_8 - Z_6) \;/\; 4C^2$$

$$G = (Z_5 - Z_4) \;/\; 2C$$

$$H = (Z_2 - Z_7) \;/\; 2C$$

Figure 11-9: Profile curvature and plan curvature measure the local terrain shape. Formulas (left) combine values surrounding a center cell with coefficients that reveal concavity or convexity in the level (plan) and downhill (profile) directions (below).

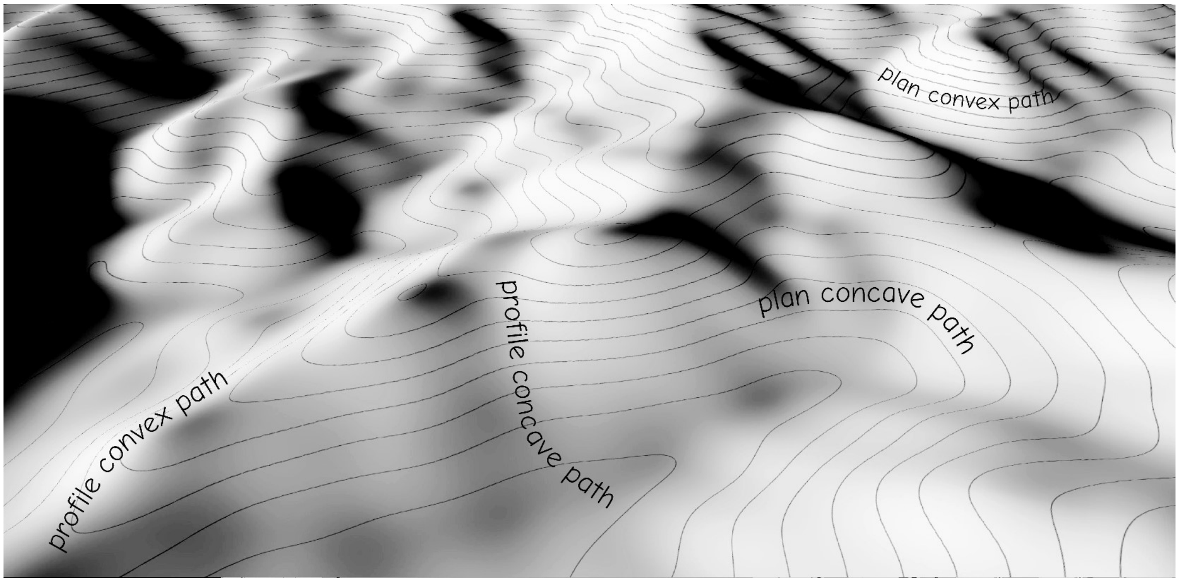


plan curvature

$$\frac{2\,(DH^2 + EG^2 - FGH)}{G^2 + H^2}$$

profile curvature

$$\frac{-2\,(DG^2 + EH^2 + FGH)}{G^2 + H^2}$$



Different softwares apply different sign conventions, sometimes making concave curvature positive, sometimes assigning them negative values. Raw values are reported in some versions, while other softwares scale curvatures over a standard range, e.g., from 0 to 100. As with most spatial analysis, the specific software implementation should be verified over known test cases.

Plan curvature is the profile shape in the local direction of level, at right angle to the steepest direction. This means plan curvature is measured at a right angle to profile curvature (Figure 11-9). Plan curvature may also be envisioned as a vertical plane slicing into the surface, and is measured in a horizontal plane. The surface traces a path on the face of the plane, and the plan curvature is a measure of the shape of that path. Concave plan curvature values are small or negative for sloping valleys or clefts, while convex plan curvature values at ridge and peak sites are large or positive.

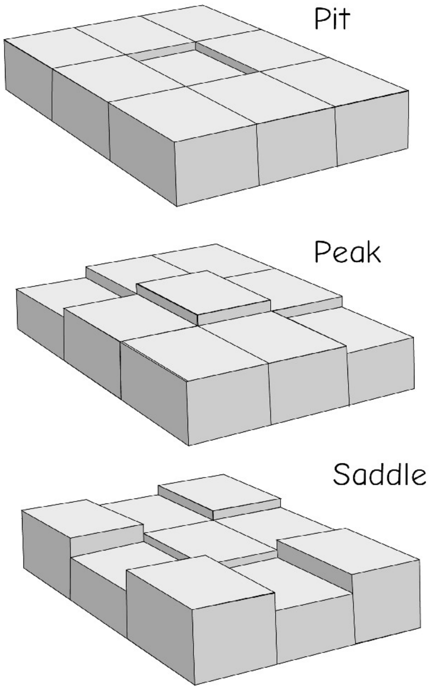


Figure 11-10: Morphometric feature types may be defined by the relative heights, and hence directional convexity, of adjacent cells (adapted from Jo Wood, 1996).

These concepts of directional terrain shape may be developed further to identify *terrain* or *morphometric features*. These are characteristic terrain elements including planes, peaks, passes, saddles, channels, ridges, shoulders, toe slopes, and pits (Figure 11-10). Each of these shapes has particular terrain attributes that often affect important spatial variables. For example, soil is thinner and water scarcer on ridges and peaks because they are convex, while materials accumulate in pits and channels.

Terrain features may be identified by observing the convexity in the plan and profile directions. For example, peaks are characterized by convex shapes in both the x and y directions, a ridge is convex in one direction but relatively flat in another, while a pit is concave in orthogonal directions (Figure 11-10). Formulas similar to those in Figure 11-9 have been developed to measure the convexity and concavity in specified orthogonal directions. The various combinations may then be applied to identify terrain features (Figure 11-11).

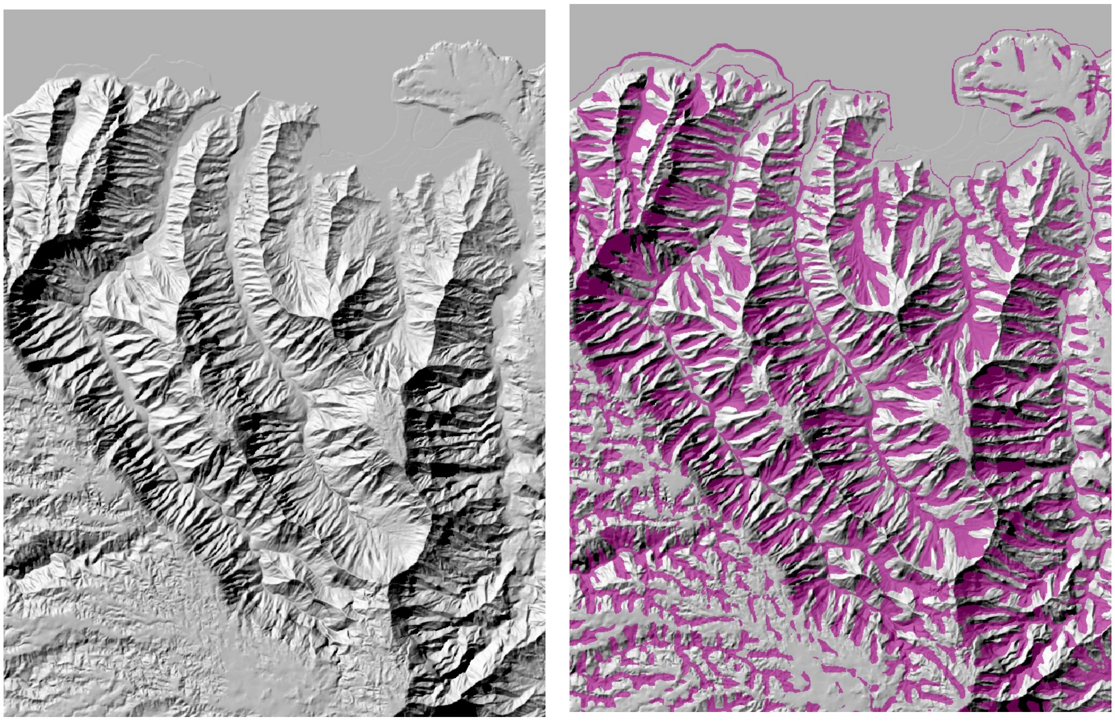


Figure 11-11: Terrain morphometery, or morphometric features may be derived from directional convexity measures. A shaded map of a mountainous area shows valleys and channels (above, left), which are identified via morphometric terrain analyses and shown as uniformly shaded areas (above, right).

Hydrologic Functions

Digital elevation models are used extensively in hydrologic analyses. Water is basic to life, commerce, and comfort, and there is a substantial investment in water resource monitoring, gathering, protection, and management. Spatial functions are applied to DEMs to yield important information on hydrology.

A *watershed* is an area that contributes flow to a point on the landscape (Figure 11-12). Watersheds may also be named basins, contributing areas, catchments, drainages, and subbasins or subcatchments. The entire uphill area that drains to any point on a landscape is the watershed for that point. Water falling anywhere in the upstream area of a watershed will pass through that point. Watersheds may be quite small. For example, the watershed may cover only a few square meters on a ridge or high slope. Local high points have watersheds of zero area because all water drains away. Watersheds may also be quite large, including continental areas that drain large rivers such as the Amazon or Mississippi Rivers. Any point in the main channel of a large river has a large upstream watershed.

The *drainage network* is the set of streams and rivers in a watershed, and it is completely contained within the watershed. As shown in Figure 11-12, the stream network often shows a dendritic pattern, with smaller watercourses branching off from larger segments as one moves upstream. The base of the drainage network is often called a *pour point* or *outlet*.

Flow direction is used in many hydrologic analyses. The true surface flow direction is the path water would take, if dumped in sufficient excess on a point so as to generate surface flow. This excess water flows in the steepest downhill direction, usually set equivalent to the local aspect.

The use of aspect to assign flow direction may be wrong, particularly in nearly flat areas and in built environments. Water flows both above and below the surface; if subsurface flow is large, ignoring it may cause errors. If soils have different permeabilities, or resistance to flow, then subsurface flow direction may be different than surface flow direction. In steep, undeveloped terrain, there is a strong downslope gravitational gradient that often dominates, and surface and subsurface flow directions are often similar, so aspect provides a reasonable approximation of overall flow direction. In flat or nearly flat terrain, soil permeability may dominate, causing different subsurface and surface flow directions. Ditches, culverts,

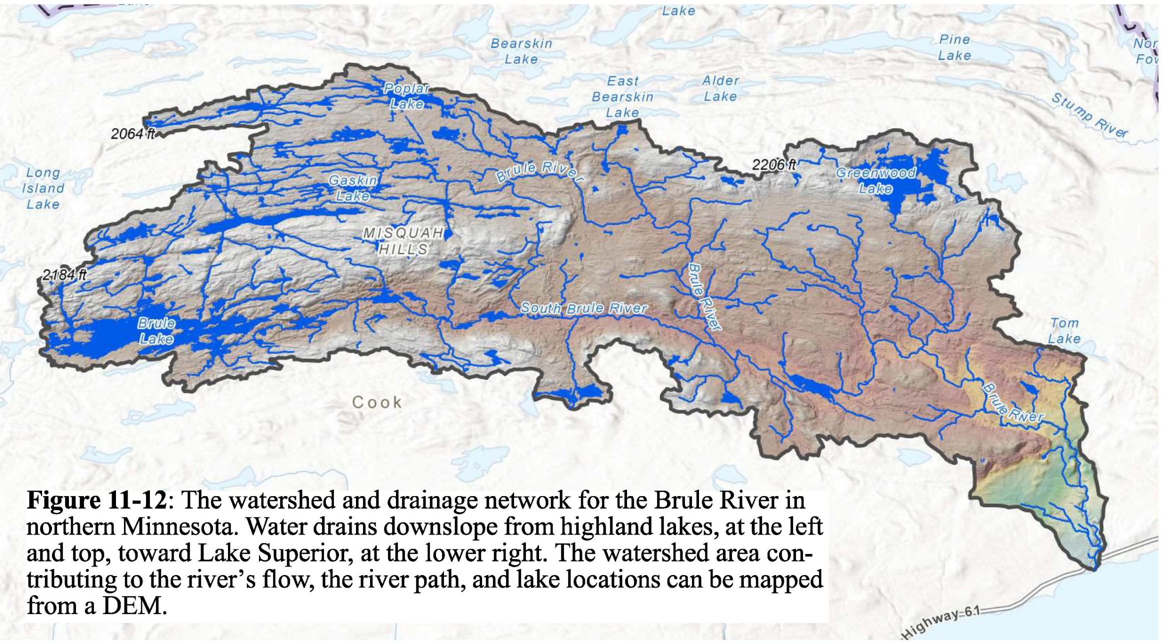


Figure 11-12: The watershed and drainage network for the Brule River in northern Minnesota. Water drains downslope from highland lakes, at the left and top, toward Lake Superior, at the lower right. The watershed area contributing to the river's flow, the river path, and lake locations can be mapped from a DEM.

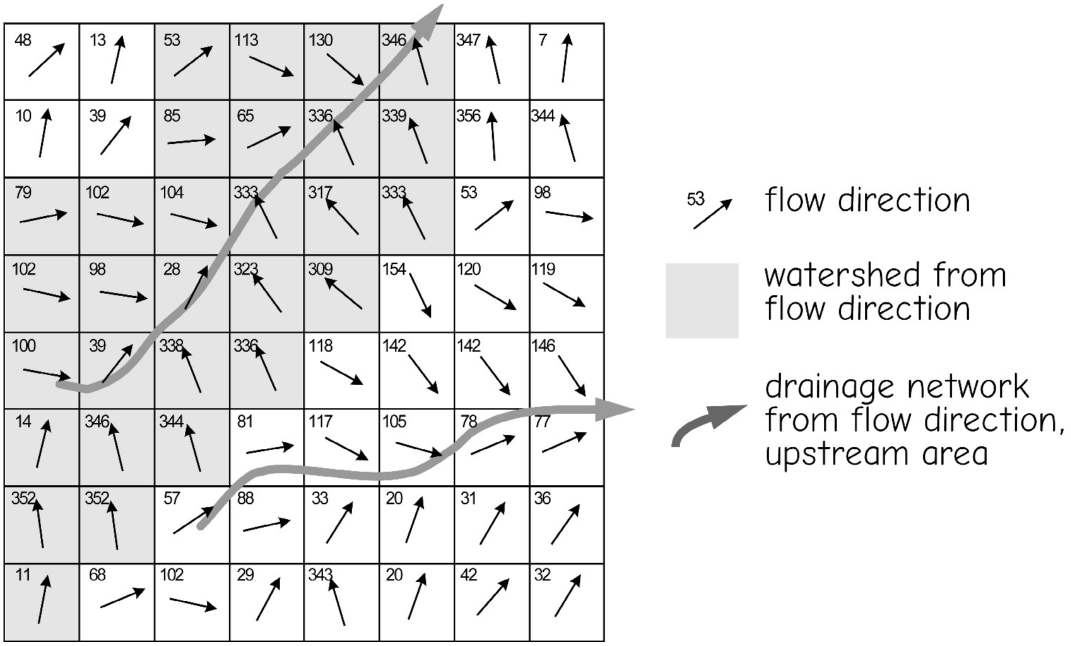


Figure 11-13: Flow direction (arrow, and number reflecting azimuth degrees), watershed, and drainage network shown for a raster grid. Elevation data are used to define the flow direction for each cell. These flow directions are then used to determine a number of important hydrologic functions.

buried storm sewers, and other built features alter flow directions in ways that aren't represented by terrain. However, subsurface drainage and built features are often based on modified flow directions that are first derived from surface shape.

Flow directions may be envisioned as an arrow from a single cell to a single adjacent cell, and stored as compass angles in a raster data layer (Figure 11-13). Acceptable values are from 0 to 360 if the angle is expressed in degrees azimuth. Alternately, flow direction can be stored as a number indicating the adjacent cell to which water flows, taking a value from 1 to 8 or some other unique identifier for each direction towards cells.

The use of a single flow direction is an incomplete representation in many instances. Cells often exhibit divergent flow, in multiple directions out of a cell to multiple adjacent cells (Figure 11-14). Flows may also be convergent, with multiple cells contributing to a cell. The most common flow direction methods provide a single direction for each cell, so divergent and some convergent flows are not represented. One solution involves recording sub-cell flow directions, but this leads to more complicated raster structures and calculations.

When the flow direction arrow from one cell does not point exactly at the adjacent cell, we may distribute the flow to more than one adjacent cell. There are various ways to distribute flows among adjacent cells. The *D8* method is common, and assigns all flow from a cell to the cell with the steepest

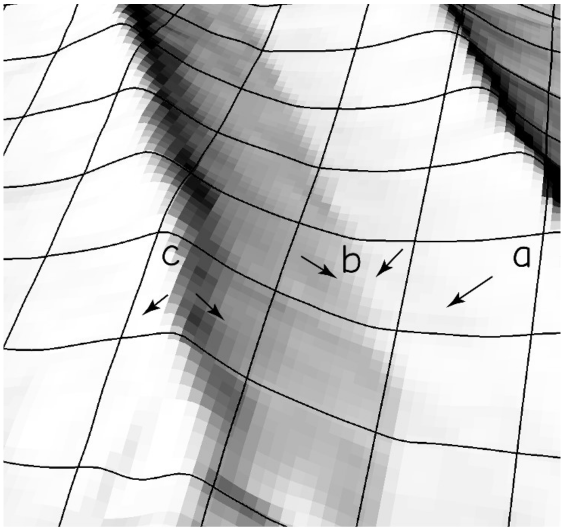


Figure 11-14: An example of simple (a), convergent (b), and divergent (c) flow.

downhill gradient (Figure 11-15, left). The D8 is simple to understand, program, and store, but is particularly poor at representing divergent flow and flow in low-gradient areas. This can cause large errors in derived measures such as upslope contributing area or soil moisture indexes, and lead to atypical drainage networks in nearly flat areas. Output flow direction rasters derived from the D8 method may be represented with only 8 codes, allowing a simple and compact data layer.

Alternative flow direction methods may assign flow to multiple cells, and hence represent some forms of divergent flow. One common method, known as *D-infinity*, distributes flow to one downslope cell when the flow direction is exactly toward the center of the cell, and otherwise assigns a portion of the flow to each of the two adjacent cells in the downslope direction (Figure 11-15, right). The split is proportional to the angles between the steepest downslope direction and the respective cell centers. This reduces the main shortcoming of the D8 method, while slightly increasing complexity.

While perhaps more accurate in many conditions, multiflow direction systems have not been widely implemented. A more common option is to use higher-resolution raster data such that raster cell size is small enough to make within-cell divergence or convergence impacts negligible

Flow accumulation area, contributing area, or *upslope area* are other important hydrologic characteristics. A flow accumulation area function is based on a flow direction surface. The flow accumulation function places a value in each cell that is the area uphill that drains to that cell.

Watersheds may be identified once a flow direction surface has been determined. Flow direction is followed "uphill" from a point, until a peak is reached. Each uphill cell may have many contributing cells, and the flow into each of these cells is also followed uphill. The uphill list is accumulated recursively until all cells contributing to the starting cell have been identified, and thus the watershed is defined.

Flow direction in flat areas is difficult to calculate and prone to error. Aspect is undefined in a truly flat region, because there is zero gradient. Flow directions in these cases may be strongly influenced by small height errors, so flow directions are sometimes

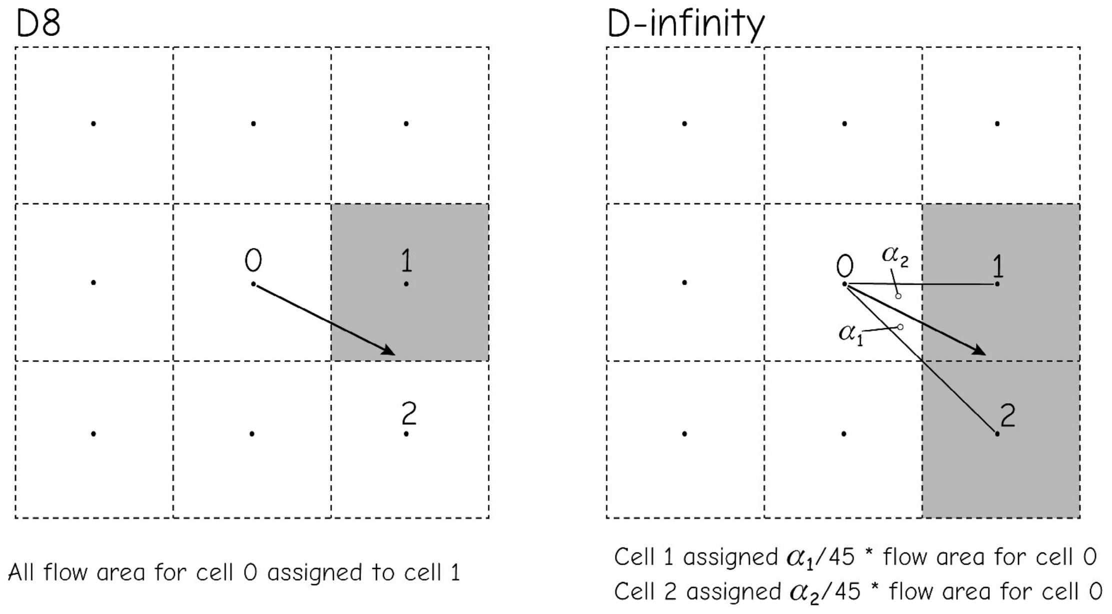


Figure 11-15: The D8 flow direction method (above left) assigns all flow to the cell center closest to the flow direction (cell 1), while the D-infinity method partitions the flow to the two cells nearest the flow direction, proportional to the flow direction angles (cells 1 and 2, above right).

manually specified, or the aspect calculated using a larger cell size or neighborhood. The neighborhood may be successively expanded until an unambiguous flow direction is defined.

Vector incision is another common method for prescribing flow direction in flat areas. A vector flowpath, e.g., a digitized stream segment, is overlain with the raster, and raster values modified downstream along the vector to specify an appropriate flow direction. The most common approach lowers elevations along the flowpath, taking care to not create a sink along or at the end of the flowpath.

A *drainage network* is the set of cells through which surface water flows. Streams, creeks, and rivers occur where flow directions converge. Thus, a flow direction may be used to produce a map of likely stream location, prior to field mapping a stream (Figure 11-13, Figure 11-16). A drainage network may be defined as any cell that has a contributing uphill area larger than some threshold. These drainage networks are only approximations, because the method does not incorporate soil texture, depth, porosity, subsoil water movement, or other properties that affect surface flow. Nonetheless, a drainage network derived from terrain data alone is often a useful first approximation. The uphill area for each cell may be calculated, and the area compared to the threshold area. The cell is marked as part of the drainage network if the area surpasses the threshold.

A drainage network may have discontinuous lines when local small dams or sink areas capture flow, where all surrounding cells point into, and none out of, a location (Figure 11-16). This may create cells immediately downhill from the sink that has a zero

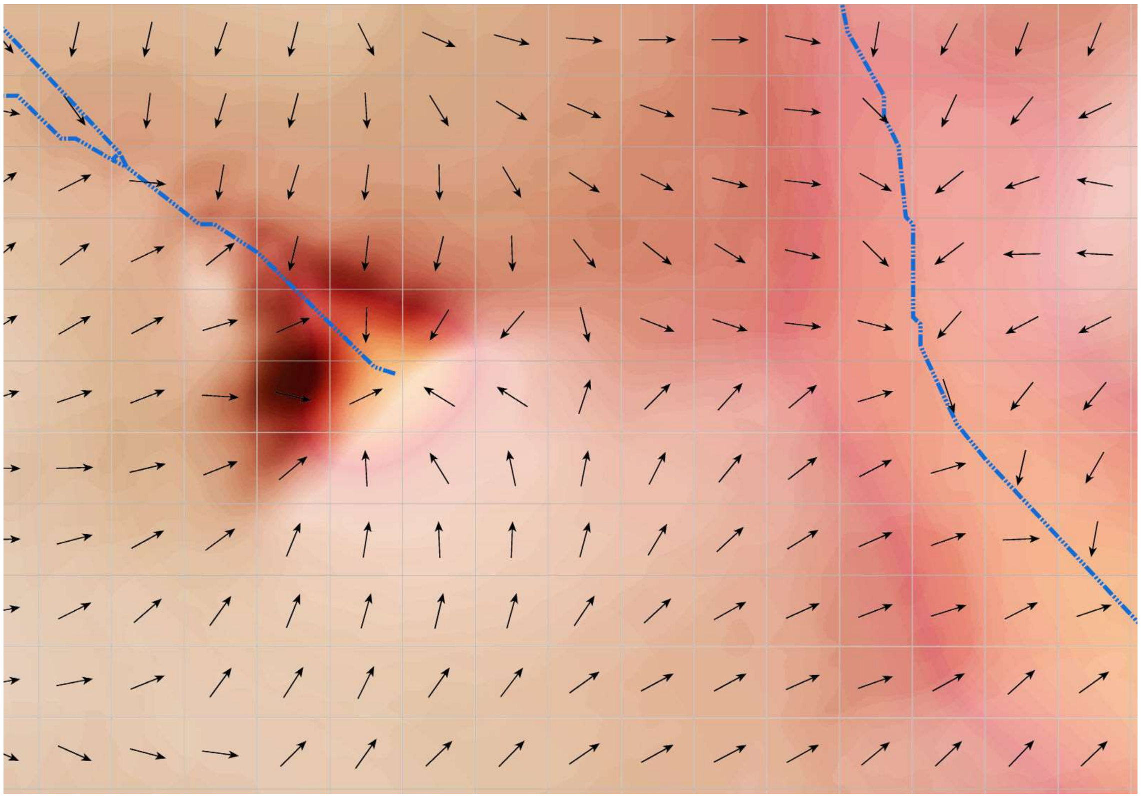


Figure 11-16: Flow direction (black arrows) calculated over a surface. The right side shows a flowline (dotted, blue in electronic edition) along a convergent zone, with values above the surface accumulation threshold. On the left of the figure is a local sink, with all flow directions pointing inward. The flowline from the upper left enters the sink and does not exit.

upstream area. The stream will end, but then may begin again further downhill. Natural sinks may be quite common in *karst* regions. where sinkholes occur on the surface due to collapsed subterranean caverns (Figure 11-17). Hydrologic sinks are also common along drainage ways in dry areas where check dams are built to reduce flash flooding or store water. Pits are also common in areas of deranged topography, for example, in the relatively flat, recently glaciated terrain. In most other areas, pits are often data artifacts and do not represent real geography. Pits represented in DEMS should be evaluated earlier during processing to determine if they are real, and how processing alters results.

Random errors in DEM elevation values often create spurious *pits* (also known as *false sinks)*. Because our technologies for creating DEMs are imperfect, DEMs often contain these pits that aren't on the Earth's surface.

Spurious pits are found in most DEMs due to small elevation errors. For example, DEM data collected with LiDAR often have a small ground footprint, and may sample small features that are above the surrounding ground level. A laser image over a recently plowed field may return spot heights for local mounds and furrows, incompletely harvested crops, and farm machinery. A log or dense shrubs in a steep-sided ravine may be misidentified as the ground surface, creating a barrier in the data that doesn't represent true conditions. Pits can be artifacts of interpolation methods that are used to fill in the grid values in unsampled locations. Post processing aims to remove these spurious readings, but they are common nonetheless.

Pits may cause problems over locally flat surfaces, often along drainage ways (Figure 11-18). Flow direction and flow accumulation functions often return errors due to spurious pits, particularly near watercourses. These low areas are shown as white patches in the figure. These apparent ponds do not exist in many landscapes, in that an erroneous pit in a stream course creates false basins.

Pits causes errors in subsequent hydrologic calculations. Drainage networks are incomplete, flow accumulation values are too low, and watersheds may be improperly identified when pits are encountered (Figure 11-19).

Figure 11-17: Examples of a discontinuous stream network, with downslope flow stopping at a sink (A), and then restarting downhill (B) when the flow accumulation threshold is again surpassed. The same kind of break is observed at a road crossing (C and D), although there is likely a culvert, not represented in the DEM.

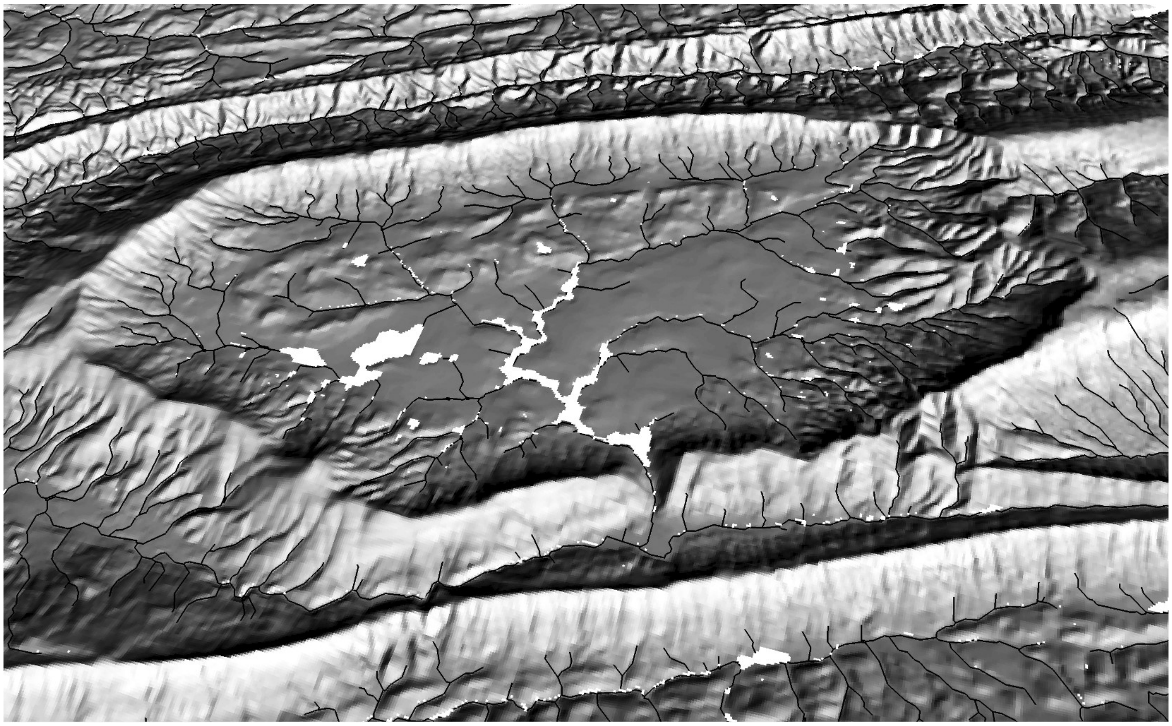


Figure 11-18: Examples of erroneous pits caused by DEM errors. The light-colored areas along drainage ways show local depressions that are artifacts of data errors, and don't exist on the landscape. Drainage networks and watersheds based on unfilled DEMs will be in error, because the flow directions based on the DEMs will be inward at all pits.

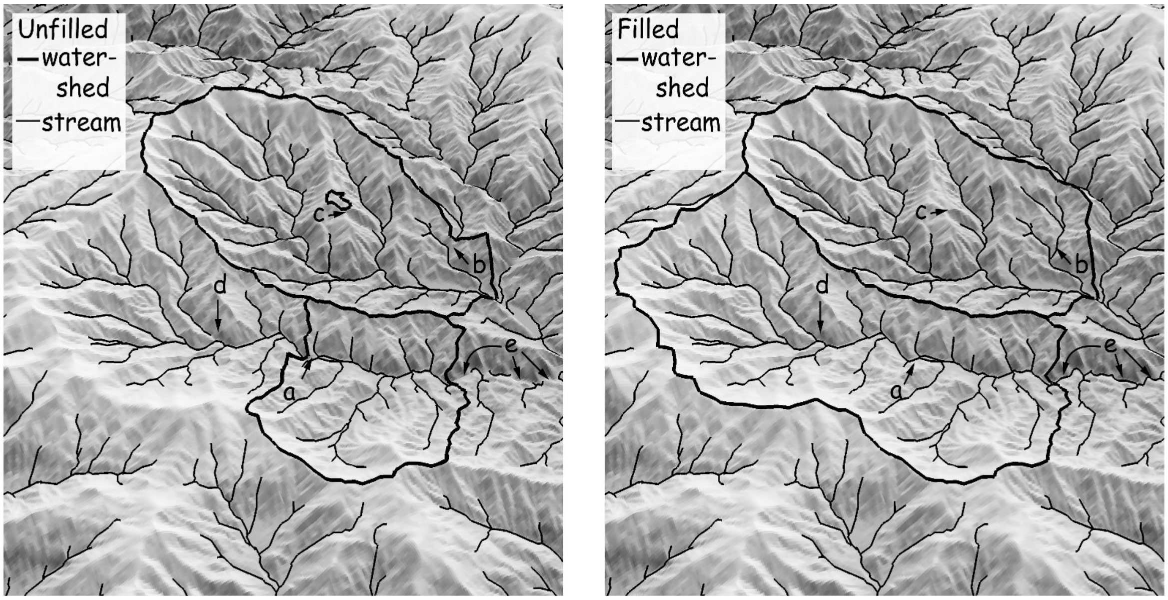


Figure 11-19: Watersheds and stream networks delineated from unconditioned DEMs (above, left) often result in missing stream segments, shown at a, b, d, and e, and incomplete watersheds (upstream from a and b, and at c).

DEMs must be "conditioned" to remove erroneous depressions (Figure 11-20). This involves pit identification, followed by either filling or downcutting downstream cells to remove the pit. A threshold is often specified above which a pit is not removed. This threshold is typically larger than common vertical errors in the data but also less than any true, "on the ground," pit depth. Known pits may be identified prior to the filling process and left unfilled. Once spurious pits are removed, further processing to identify watersheds and drainage networks may proceed.

The pit may substantially alter the DEM, and the scope of alteration may depend on the method (Figure 11-21). A fill process raises the values of a local depression until all cell values are at least equal to the value at the local "rim" or edge of the depression (Figure 11-21, center). This may create a flat surface, with no unambiguous drainage direction, so some variants of the fill process add a small slope over large fill areas to ensure drainage toward a downhill direction. Pits may also be removed through a breaching process (Figure 11-21, bottom), in which cells along a steepest gradient are lowered, searching a specified surrounding area to identify the steepest downhill path.

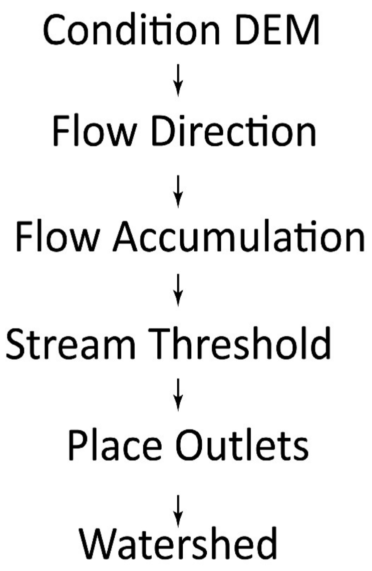


Figure 11-20: Steps in a watershed delineation.

As shown in Figure 11-21, breaching may sometimes better reflect the true drainage pathways rather than fills, and may result in more "natural" landscapes. It often depends on the nature of the depression, whether it is due to a spurious, small, isolated low elevation value (fill usually preferred for conditioning), or a narrow, high, linear feature, often built and with a culvert or other subsurface drainage way (breach usually preferred for conditioning). Unfortunately, many GIS softwares do not provide a breach function, even though breaching is increasingly useful for high-resolution DEMs based on LiDAR over urban or built-up areas.

Drainage and watershed geography inferred from terrain analysis depend substantially on the algorithm used, particularly for flow direction, so care should be taken in identifying the methods and thresholds that give sufficiently accurate results for the intended tasks. Many softwares only provide depression filling, and D8 flow direction, and often result in erroneous flowpaths in flat or near-flat terrains. The broadest range of general hydrologic and general terrain analysis tools are currently provided by Whitebox GAT, developed and maintained by John Lindsay at the University of Guelph.

To review, the steps for identifying a watershed from a DEM is shown inFigure 11-22. DEMs are conditioned as needed, and then the flow direction, accumulation, stream threshold, and watershed boundaries calculated. Different conditioning and flow accumulation methods may result in slightly different stream locations and, in some cases, watershed boundaries.

Several other hydrologic indices have been developed to identify locally convergent or divergent terrain positions, or terrain morphometry related to hydrography. These indexes are used in many subsequent topographic and hydrologic analyses, such as predicting plant community composition or growth, erosion modeling, or estimating the rainfall required to saturate an area and predict the likelihood and intensity of flooding.

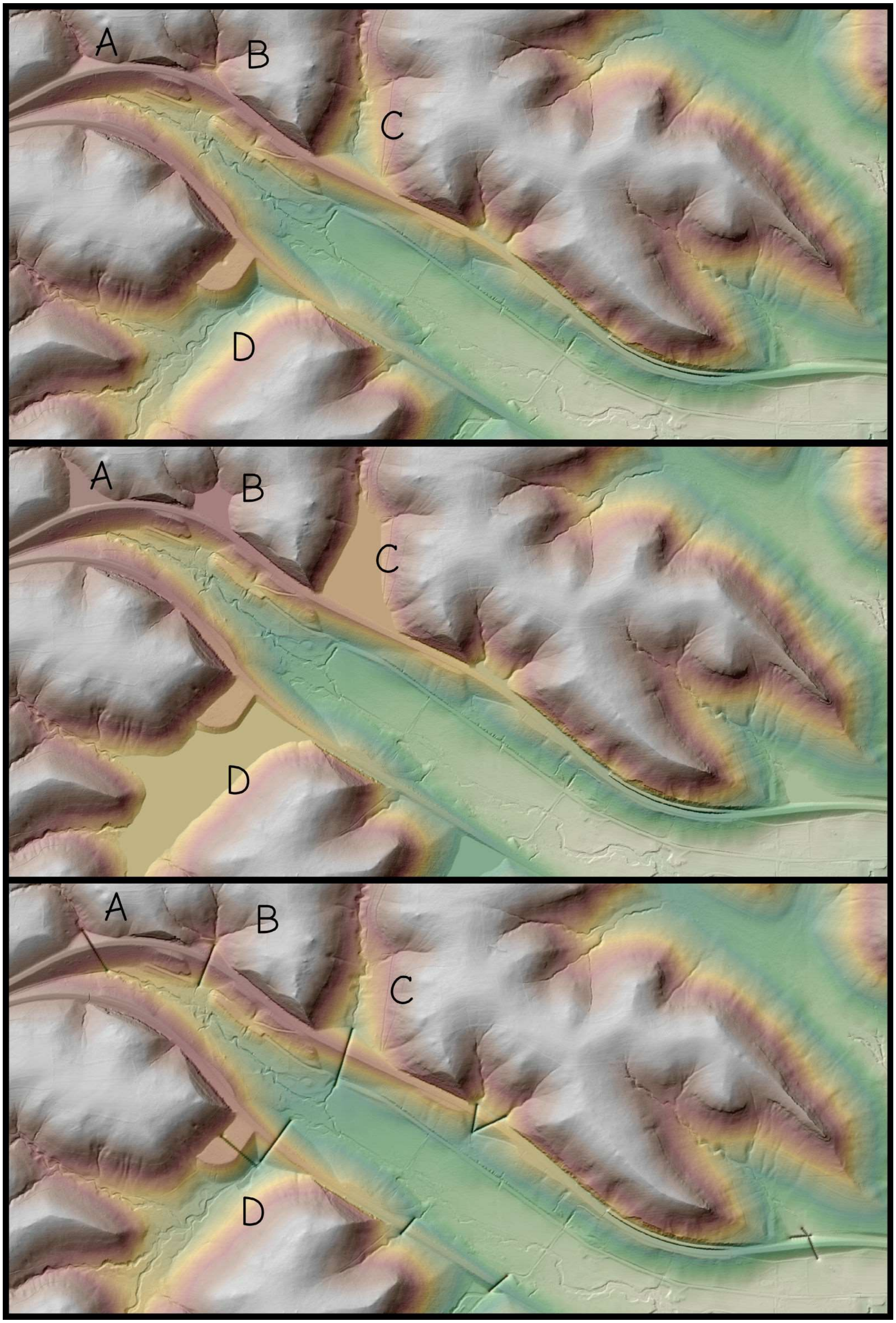


Figure 11-21: A DEM (top) with large sinks at A through D as a result of highway berms, with several smaller sinks in other locations. Sinks are removed by either a fill process (middle), or breaching (bottom). Breaching results in an output surface that is more accurate for most applications.

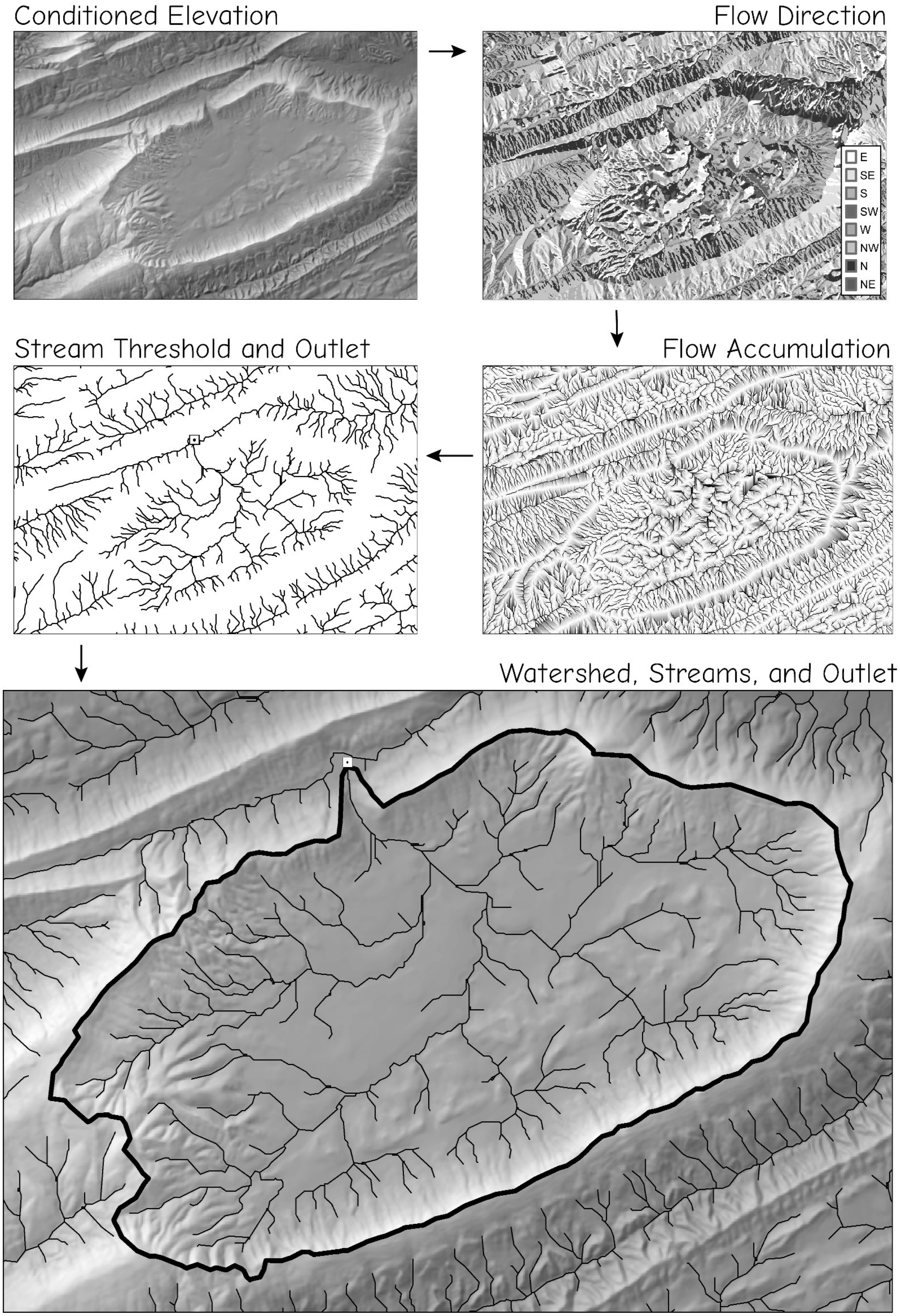


Figure 11-22: An example of the steps required to create watershed and drainage network features from a digital elevation model.

The *specific catchment area* (SCA) is defined as the total area draining to a point relative to drainage width, in raster data sets calculated as

$$SCA = AREA/C \qquad \textbf{(11.5)}$$

where AREA is the accumulated surface area upstream from a point, and C is the raster cell dimension. *Stream power index* (SPI) is defined as:

$$SPI = SCA * \tan(b) \qquad (11.6)$$

where b is the slope at a point, and SCA is as defined above. SPI is used to identify the potential erosion at a point, which depends both on the upstream area and hence ability to accumulate water, and the local slope, which drives the erosive energy in water flow.

Perhaps the most commonly applied wetness index is calculated by:

$$w = \ln\left(\frac{SCA}{\tan\beta}\right) \qquad \textbf{(11.7)}$$

where w is the wetness index at a cell, SCA is the specific catchment area, and β is the slope at the cell. This index has been shown to effectively represent the increased soil wetness due to large upslope areas and low slopes, particularly when combined with plan curvature and profile curvature measurements. These factors sort terrain along ridge-to-stream and convex-to-concave gradients.

There are many other topographic indices, e.g., for estimating total solar radiation, surface air drainage, or surface roughness. These and others are described in the references at the end of this chapter.

Contour Lines

Contour lines, or topographic contours, are connected lines of uniform elevation that run at right angles to the local slope. Contour lines are a common feature on many map series; for example, they are depicted on the USGS 1:24,000 scale nationwide series, and Britain's 1:50,000 Ordinance Survey maps. The shape and density of contour lines provide detailed information on terrain height and shape in a two-dimensional map, without the need for continuous tone shading. Both color and continuous tone printing were important limitations for past cartographers. Contour lines could be easily drawn with simple drafting tools. Although continuous tone printing is much less expensive today, contours will remain common as they have entered the culture of map making and map reading.

Several rapid, efficient methods have been developed for calculating contours, either from points or from grid data (Figure 11-23). Early contour maps and DEMs were developed from height measurements at a set of points. While useful, these points did not provide clear depictions of elevation. Contour lines of fixed values were interpolated

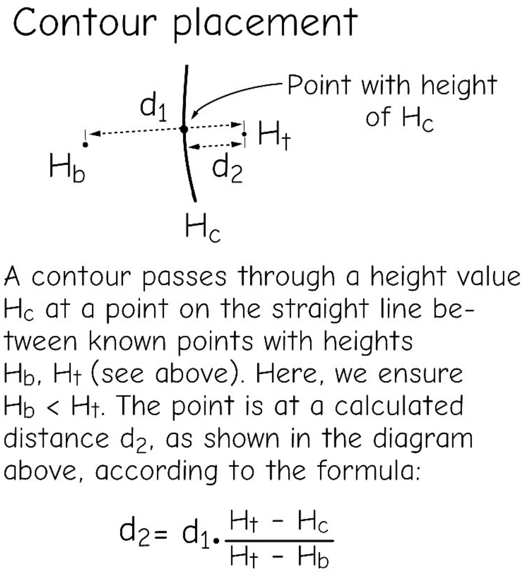


Figure 11-23: Contour line locations are often estimated from point height locations, as a linear proportion of the height and distance differences between points.

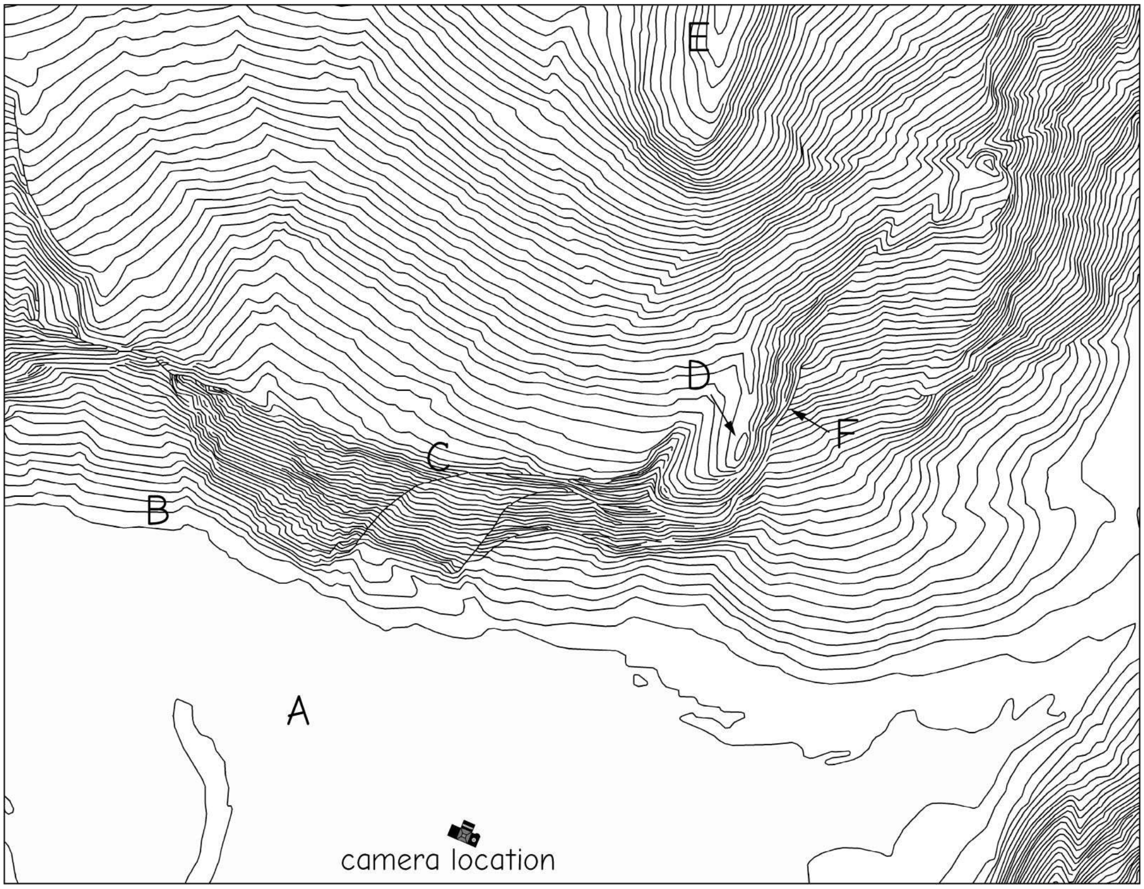


Figure 11-24: Contours may depict terrain features succinctly, as shown in this contour map, looking vertically downward on the area approximately shown in the photo. The image includes the Royal Arches and North Dome, from Yosemite Valley, California. Changes in slope, cliffs, overhangs, and peaks are all succinctly represented by contour lines.

linearly between nearest measurement points, as shown in Figure 11-23. Later measurement methods either identified contour lines directly from stereopairs (see Chapter 6), or derived them from mechanically or electronically produced rasters. Raster to contour generation also typically follows a linear interpolation. For a raster, appropriate adjacent cell centers are selected, and contour values interpolated as illustrated in Figure 11-23.

Contour lines are typically created at fixed height intervals, for example, every 30 m (100 ft) from a base height (Figure 11-24). Because each line represents a fixed elevation above or below adjacent lines, the density of contour lines indicates terrain steepness. Point A in Figure 11-24 falls in a flat area (the foreground of the photo, at bottom), where elevation does not change much, and there are few contour lines. Steep areas and cliffs are depicted by an increase

in contour density, as shown at point B, with changes in steepness depicted by changes in density (above and below point C). Peaks, such as the top of Washington's column, D, and North Dome, E, appear as concentric rings. Note that contours may succinctly represent complex terrain structures, such as the curving arches in the center of the photograph, and shown below point C, and the overhanging cliff, to the left of point F.

Profile Plots

Profile plots are another common derivative of elevation data. These plots sample elevation along a linear *profile path*, and display elevation against distance in a graph (Figure 11-25). Elevation is typically plotted on the y axis, and horizontal distance on the x axis These profile plots are helpful in visualizing elevation change, slope, and cumulative travel distance along the specific profile path. Profile plots are common on the edges of maps, particularly maps of off-road, bicycle, or cross-country routes.

Profile plots often have some level of vertical exaggeration because horizontal distances are usually much larger than elevation gain. Vertical exaggeration is a scaling factor applied to the elevation data when shown on the graph. For example, Figure 11-25 shows a square graph that depicts approximately 31 km across the Earth's surface. The vertical elevation axis spans approximately 2.5 km over the same dimensions on the graph. This is a vertical exaggeration of approximately 12 (from 31/2.5).

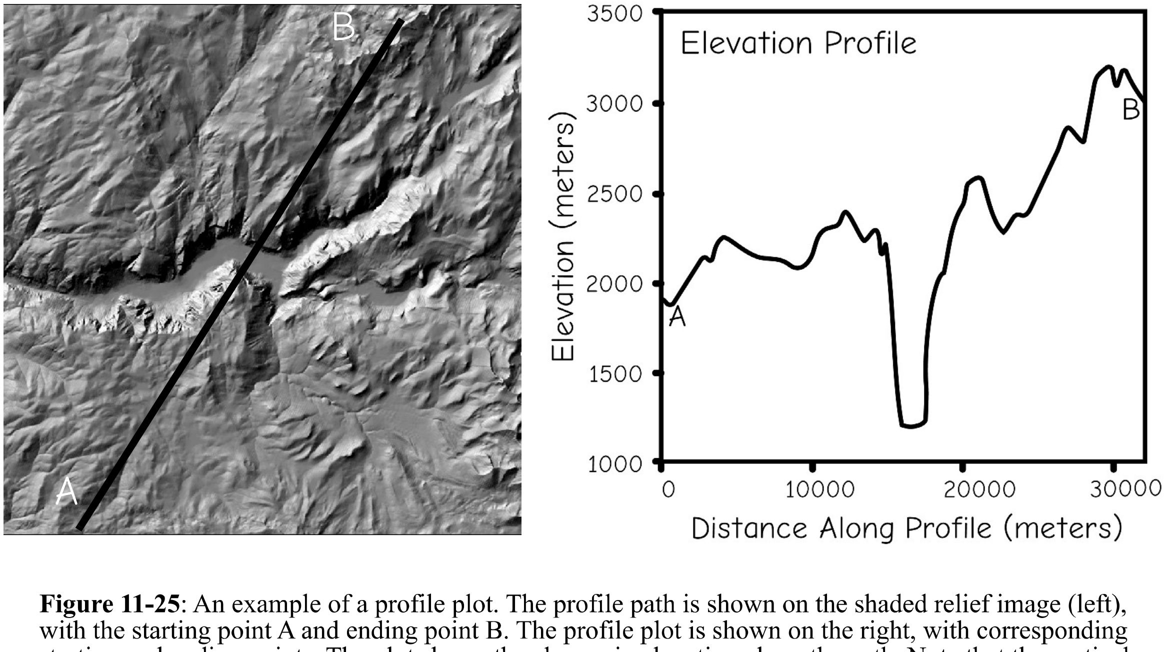


Figure 11-25: An example of a profile plot. The profile path is shown on the shaded relief image (left), with the starting point A and ending point B. The profile plot is shown on the right, with corresponding starting and ending points. The plot shows the change in elevation along the path. Note that the vertical exaggeration here is approximately 9 to 1.

Viewsheds

The *viewshed* for a point is the collection of areas visible from that point. Views from many locations are blocked by terrain. Elevations will hide points if the elevations are higher than the line of sight between the viewing point and target point (Figure 11-26).

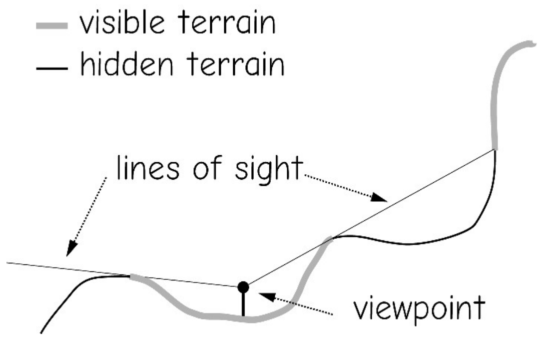


Figure 11-26: Mechanics of defining a viewshed.

Viewsheds and visibility analyses are quite important in many instances. High-voltage power lines or cell towers are often placed after careful consideration of their visibility, because most people are averse to viewing them. Communications antennas, large industrial complexes, and roads are often located at least partly based on their visibility, and viewsheds are specifically managed for many parks and scenic areas.

A viewshed is calculated based on cell-to-cell intervisibility. A line of sight is drawn between the view cell and a potentially visible target cell (Figure 11-26). The elevation of this line of sight is calculated for every intervening cell. If the slope to a target cell is less than the slope to a cell closer to the viewpoint along the line of sight, then the target cell is not visible from the viewpoint. Specialized algorithms have been developed to substantially reduce the time required to calculate viewsheds, but in concept, lines of sight are drawn from each viewpoint to each cell in the digital elevation data. If there is no intervening terrain, the cell is classified as visible. The classification identifies areas that are visible and areas that are hidden (Figure 11-27). Viewsheds for line or area features are the accumulated viewsheds from all the cells in those features.

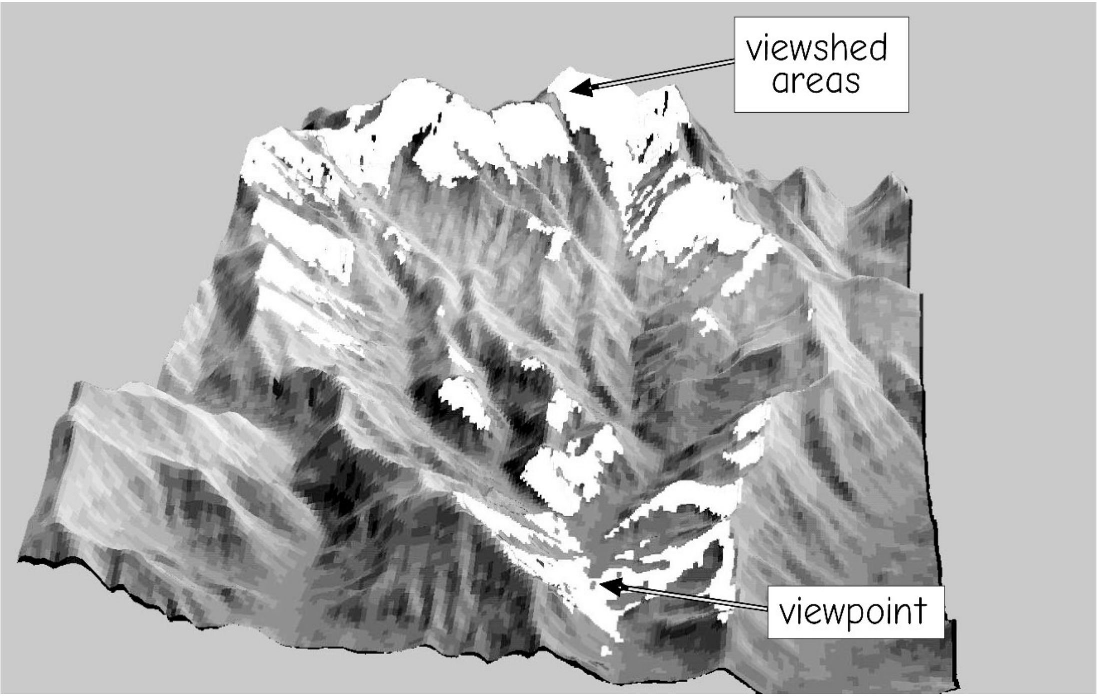


Figure 11-27: An example of a viewpoint, and corresponding viewshed.

Shaded Relief Maps

A *shaded relief map*, also often referred to as a *hillshade map*, is a depiction of the brightness of terrain reflections given a terrain surface and sun location. Although shaded relief maps are rarely used in analyses, they are among the most effective ways to communicate the shape and structure of terrain features, and many maps include relief shading (Figure 11-28).

Shaded relief maps are developed from digital elevation data and models of light reflectance. An artificial sun is "positioned" at a location in the sky and light rays projected onto the surface depicted by the elevation data. Light is modeled that strikes a surface either as a direct beam, from the sun to the surface, or from background "diffuse" sunlight. Diffuse light is scattered by the atmosphere, and illuminates "shaded" areas, although the illumination is typically much less than that from direct beam.

The brightness of a cell depends on the local incidence angle, the angle between the incoming light ray and the surface normal, shown as θ in Figure 11-29. The surface normal is defined as a line perpendicular to the local surface. Direct beam sunlight striking

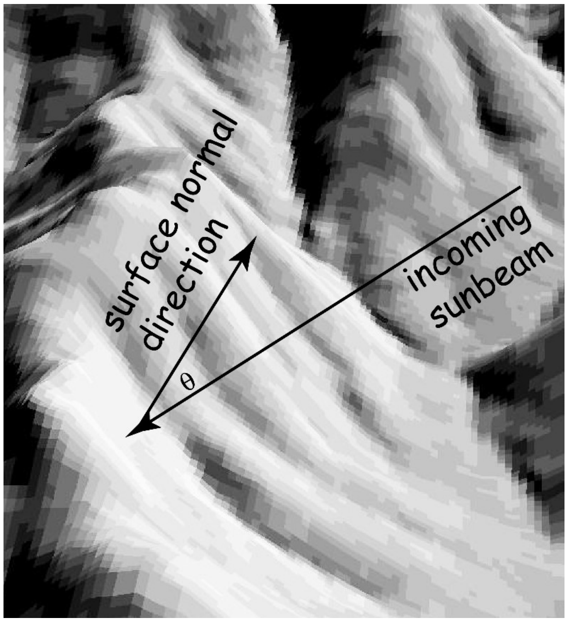


Figure 11-29: Hillshade maps show reflectance as a function of the angle, θ, between sunbeams and surface normals.

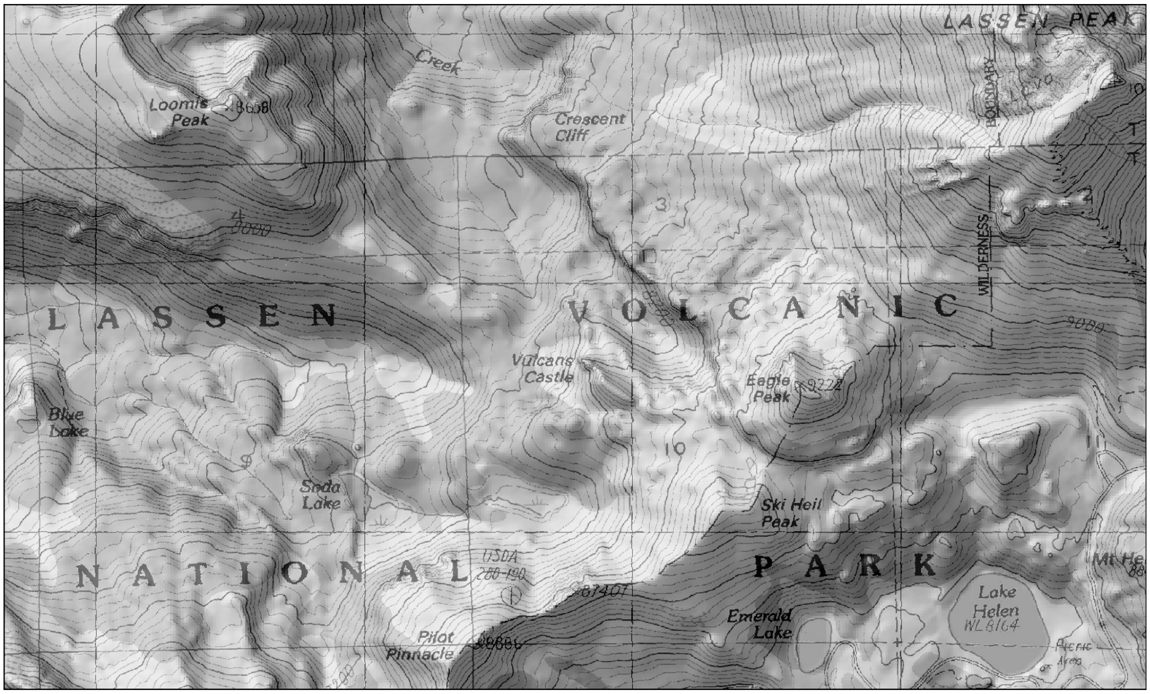


Figure 11-28: Relief shading is often added as a background "under" other mapped data to provide a sense of terrain shape and steepness. This shading provides a three-dimensional perspective for a mapped area, as demonstrated in this relief shading of a U.S. Geological Survey 1:24,000 scale quandrangle map.

the surface at a right angle (θ = 0) provides the brightest return, and hence appears light. As θ increases, the angle between the direct beam and the ground surface deviates from perpendicular, and the brightness decreases. Diffuse sunlight alone provides a relatively weak return, and hence appears dark. Combinations of direct and diffuse light result in a range of gray shades, and this range depends on the terrain slope and angle relative to the sun's location. Hence, subtle variations in terrain are visible on shaded relief maps.

Calculating a shaded relief surface requires specifying the sun's position, usually via the solar zenith angle, measured from vertical down to the sun's location, and the solar azimuth angle, measured from north clockwise to the sun's position (Figure 11-30). Local slope and surface azimuth define a surface normal direction. An angle may be defined between the solar direction and the surface normal direction, shown as θ in Figure 11-30. As noted earlier, the amount of reflected energy decreases as θ increases,

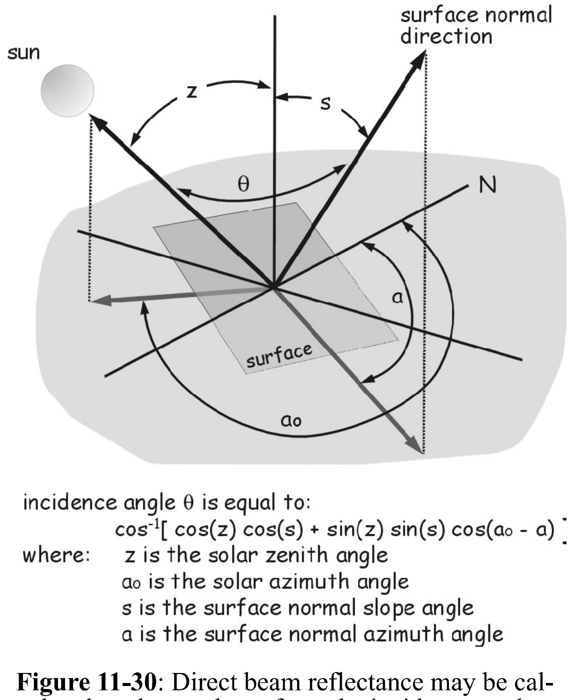


Figure 11-30: Direct beam reflectance may be calculated as shown above from the incidence angle, θ, between the incoming sunbeam and the local surface normal. The surface normal is defined by a line perpendicular to the local surface plane.

and this may be shown as various shades of grey in a hillshade surface.

A shaded relief map also requires a calculation of visibility, often prior to calculating the reflectances. Visibility to the sun is determined; if a cell is visible from the sun, the slope and aspect values are used to assign the cell brightness.

Terrain Analysis Software

Terrain analysis and DEM data management and analysis are important enough to be included in most general-purpose GIS packages, including ArcGIS, GRASS, ERDAS, Idrisi, and Manifold. While they support the most common set of terrain and hydrologic analyses, none of these packages includes the broadest range of terrain processing and analysis functions. Specialized analyses are often performed using software with a specific focus on terrain analysis. These include the Whitebox GAT, from the University of Guelph, and commercial tools, such as the Watershed Modeling System (WMS) by the Scientific Software Group.

Whitebox GAT contains what is likely the most comprehensive set of terrain analysis functions in a freely available package. Support is particularly strong for hydrologic surface and stream link processing and analysis, with functions for calculating various flow direction, accumulation, and watershed delineation methods typically not supported by other packages. Basic terrain modification, LiDAR data input and processing, and general raster GIS functions are also supported.

Landserf is a package with particularly strong support for terrain shape and geomorphological analysis, in addition to a strong focus on surface visualization. Multiple methods of calculating and combining first- and second-order terrain gradients are supported, as well as basic elevation data conversion and processing. Landserf is written in Java, and hence available across the widest range of operating systems.

ArcHydro is a set of hydrologic analysis tools written as an extension to ArcGIS. It supports a fairly complete set of hydrologic and watershed delineation functions.

There are many other packages available, including RiverTools, TAUDEM, Surfer, TAPES, and MicroDEM, which provide various specialized capabilities, and may be worth investigating for users interested in terrain and hydrologic analysis.

Summary

Terrain analyses are commonly performed within the framework of a GIS. These analyses are important because terrain governs where and how much water will accumulate on the landscape, how much sunlight a site receives, and the visibility of human activities.

Slope and aspect are two of the most used terrain variables. Both are commonly calculated via trigonometric functions applied in a moving window to a raster DEM. Several kernels have been developed to calculate changes of elevation in x and y directions, and these component gradients are combined to calculate slope and aspect.

Profile curvature and plan curvature are two other important terrain analysis functions. These functions measure the relative convexity or concavity in the terrain, relative to the downslope direction for profile curvature and the cross-slope direction for plan curvature.

Terrain analyses are also used to develop and apply hydrologic functions and models. Watershed boundaries, flow directions, flowpaths, and drainage networks may all be defined from digital elevation data.

Viewsheds are another commonly applied terrain analysis function. Intervisibility may be computed from any location on a DEM. A line of sight may be drawn from any point to any other point, and if there is no intervening terrain, then the two points are intervisible. Viewsheds are often used to analyze the visibility of landscape alterations or additions, for example, when siting new roads, powerlines, or large buildings.

Finally, relief shading is another common use of terrain data. A shaded relief map is among the most effective ways to depict terrain. Terrain shading is often derived from DEMs and depicted on maps.

Suggested Reading

Ali, G., Birkel, C., Tetzlaff, D., Soulsby, C., McDonnell, J.J., Tarolli, P. (2014). A comparison of wetness indices for the prediction of observed connected saturated areas under contrasting conditions. *Earth Surface Process and Landforms*, 39:399–413.

Ayeni, O.O. (1982). Optimum sampling for digital terrain models. *Photogrammetric Engineering and Remote Sensing*, 48:1687–1694.

Band, L.E. (1986). Topographic partition of watersheds with digital elevation models. *Water Resources Research*, 22:15–24.

Baral, D.J., Gupta, R.P. (1997). Integration of satellite sensor data with DEM for the study of snow cover distribution and depletion patterns. *International Journal of Remote Sensing*, 18:3889–3894.

Berry, J.K. (1986). A mathematical structure for analyzing maps. *Environmental Management*, 11:317–325.

Berry, J.K. (1987). Fundamental operations in computer-assisted mapping. *International Journal of Geographic Information Systems*, 1:119–136.

Beven, K.J., Kirby, M.J. (1979). A physically-based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, 24:43–69.

Bolstad, P.V., Stowe,T. (1994). An evaluation of DEM accuracy: elevation, slope and aspect. *Photogrammetric Engineering and Remote Sensing*, 60:1327–1332.

Bonham-Carter, G.F., (1996). *Geographic Information Systems for Geoscientists: Modelling with GIS*. Ottawa: Pergamon.

Burrough, P.A., McDonnell, R.A. (1998). *Principles of Geographical Information Systems* (2nd ed.). New York: Oxford University Press.

Collins, S.H., Moon, G.C. (1981). Algorithms for dense digital terrain models. *Photogrammetric Engineering and Remote Sensing*, 47:71–76.

DeFloriani, L., Magillo, P. (1994). Visibility algorithms on triangulated digital terrain models. *International Journal of Geographical Information Systems*, 8:13–41.

Dozier, J., Frew, J. (1990). Rapid calculation of terrain parameters for radiation modeling from digital elevation data. *IEEE Transactions on Geoscience and Remote Sensing*, 28:963–969.

Dubayah, R., Rich, P.M. (1995). Topographic solar radiation models for GIS. *International Journal of Geographical Information Systems*, 9:405–419.

Fisher, P.F. (1996). Reconsideration of the viewshed function in terrain modeling. *Geographical Systems*, 3:33–58.

Flint, A.L., Childs, S.W. (1987). Calculation of solar radiation in mountainous terrain. *Agricultural and Forest Meteorology*, 40:233–249.

Hengl, T., Reuter, H.I., Eds. (2009). *Geomorphometry: Concepts, Software, Applications*. Amsterdam: Elsevier.

Hodgson, M.E. (1995). What cell size does the computed slope/aspect angle represent? *Photogrammetric Engineering and Remote Sensing*, 61:513–517.

Horn, B.K. (1981). Hill shading and the reflectance map. *IEEE Proceedings on Geosciences*, 69:14–47.

Hutchinson, M.F. (1989). A new procedure for gridding elevation and stream line data with automatic removal of spurious pits. *Journal of Hydrology*, 106:211–232.

Hutchinson, M.F. (1993). Development of a continent-wide DEM with applications to terrain and climate analysis, In: M.F. Goodchild et al. (Eds.). *Environmental Modeling with GIS*. New York: Oxford University Press.

Jain, M.K., Kothyari, U.C., & Ranga, R.K.G. (2004). A GIS based distributed rainfall-runoff model. *Journal of Hydrology*, 299:105–122.

Jenson, S.K. (1991). Applications of hydrologic information automatically extracted from digital elevation models. *Hydrologic Processes*, 5:31–44.

Jenson, S.K., Domingue, J.O. (1988). Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric Engineering and Remote Sensing*, 54:1593–1600.

Jones, N.L., Wright, S.G., Maidment, D.R. (1990). Watershed delineation with triangle-based terrain models. *Journal of Hydraulic Engineering*, 116:1232–1251.

Julian, J.P., Elmore, A.J., Guinn, S.M. (2012). Channel head locations in forested watersheds across the mid-Atlantic United States: A physiographic analysis. *Geomorphology*, 177/78:194–203.

Lindsay, J.B. (2005). The terrain analysis system: a tool for hydro-geomorphic applications. *Hydrological Processes*, 19:1123–1130.

Lindsay J.B. (2016). The practice of DEM stream burning revisited. *Earth Surface Processes and Landforms*, 41(5): 658-668. DOI: 10.1002/esp.3888.

Louhaichi, M., Borman, M.M., Johnson, A.L., Johnson, D.E. (2003). Creating low-cost high-resolution digital elevation models. *Journal of Range Management*, 56:92–96.

Martz, L.W., Garbrecht, J. (1998). The treatment of flat areas and depressions in automated drainage analysis of raster digital elevation models. *Hydrological Processes*, 12:843–856.

Maune, D.F. (Ed.). (2007). *Digital Elevation Model Technologies and Applications: The DEM User's Manual* (2nd ed.). Bethesda: American Society of Photogrammetry and Remote Sensing.

Moore, I.D., Grayson, R.B. (1991). Terrain-based catchment partitioning and runoff prediction using vector elevation data. *Water Resources Research*, 27:1177–1191.

Moore, I.D., Turner, A, Jenson, S., Band, L. (1993). GIS and land surface-subsurface process modelling, In M.F. Goodchild et al. (Eds.), *Environmental Modeling with GIS*. New York: Oxford University Press.

Skidmore, A.K. (1989). A comparison of techniques for calculating gradient and aspect from a gridded digital elevation model. *International Journal of Geographical Information System*s, 3:323–334.

Southee, F.M., Treitz, P.M., Scott, N.A. (2012). Application of LiDAR terrain surfaces for soil moisture modeling. *Photogrammetric Engineering and Remote Sensing*, 12:1241–1251.

Strahler, A.N. (1957). Quantitative analysis of watershed geomorphology, *Transactions of the American Geophysical Union*, 8:913–920.

Tarboton, D.G., Bras, R.L., Rodriguez-Iturbe, I. (1992). A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resources Research*, 33:309–319.

Tomlin, C.D. (1990). *Geographic Information Systems and Cartographic Modeling*. Upper Saddle River: Prentice-Hall.

Wilson, J.P. (2012). Digital terrain modeling. *Geomorphology*, 137:107–121.

Wilson, J., Gallant, J. (Eds.). (2000). *Terrain Analysis: Principles and Applications*. New York: Wiley.

Wood, J. (1996). The geomorphological characterization of digital elevation models. Ph.D. thesis, University of Leicester, UK.

Wood, R., Sivapalan, M., Robinson, J. (1997). Modeling the spatial variability of surface runoff using a topographic index. *Water Resources Research*, 33:1061–1073.

Zevenbergen, L.W., Thorne, C.R. (1987). Quantative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12:47–56.

Zhou, Q., Liu, X. (2004). Analysis of errors of derived slope and aspect related to DEM data properties. *Computers & Geosciences,* 30:369–378.

Ziadat, F.M. (2005). Analyzing digital terrain attributes to predict soil attributes for a relatively large area. S*oil Science Society of America Journal*, 69:1590–1599.

Study Questions

11.1 - What are digital elevation models, and why are they used so often in spatial analyses?

11.2 - How are digital elevation data created?

11.3 - Write the definition of slope and aspect, and the mathematical formulas used to derive them from digital elevation data.

11.4 - Calculate dZ/dx and dZ/dy for the following 3 x 3 windows. Elevations and the cell dimension are in meters.

windows | 4-nearest cell | 3rd-order finite difference

a)

110	113	118
112	114	119
111	117	121

←10→

4-nearest cell: dZ/dx = dZ/dy=

3rd-order finite difference: dZ/dx = dZ/dy =

b)

67	63	62
65	64	64
70	68	66

4-nearest cell: dZ/dx = dZ/dy=

3rd-order finite difference: dZ/dx = dZ/dy =

c)

18	23	17
21	24	19
20	22	18

4-nearest cell: dZ/dx = dZ/dy=

3rd-order finite difference: dZ/dx = dZ/dy =

11.5 - Calculate dZ/dx and dZ/dy for the following 3 x 3 windows. Elevations and the cell dimension are in meters.

windows | 4-nearest cell | 3rd-order finite difference

a)

108	112	115
119	116	118
113	118	119

←10→

4-nearest cell: dZ/dx =

dZ/dy=

3rd-order finite difference: dZ/dx =

dZ/dy =

b)

68	63	61
69	67	66
70	71	72

4-nearest cell: dZ/dx =

dZ/dy=

3rd-order finite difference: dZ/dx =

dZ/dy =

c)

15	19	18
19	20	19
21	23	24

4-nearest cell: dZ/dx =

dZ/dy=

3rd-order finite difference: dZ/dx =

dZ/dy =

11.6 - Calculate the slope and aspect for the underlined cell values, using the four nearest cell method.

712	709	707	703	704
710	<u>706</u>	704	700	702
708	705	705	<u>697</u>	700
711	<u>709</u>	705	696	694
714	712	708	703	698

←10→

11.7 - Calculate the slope and aspect for the underlined cell values, using the four nearest cell method.

712	709	707	703	704
710	706	704	a) 700	702
708	705	b) 705	697	700
711	709	705	c) 696	694
714	712	708	703	698

←20→

11.8 - Calculate the slope and aspect for the underlined cell values, using the third-order finite difference method.

712	709	707	703	704
710	706	704	700	702
708	705	705	697	700
711	709	705	696	694
714	712	708	703	698

←10→

11.9 - Calculate the slope and aspect for the underlined cell values, using the third-order finite difference method.

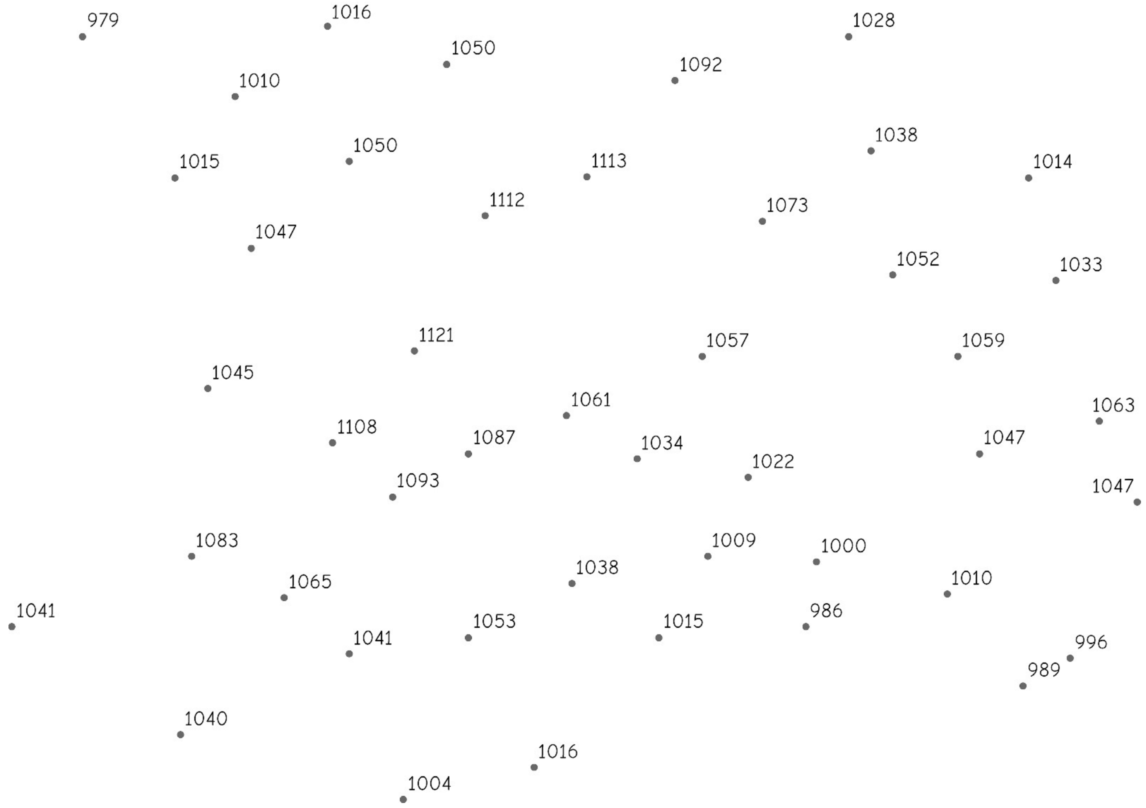
712	709	707	703	704
710	706	704	a) 700	702
708	705	b) 705	697	700
711	709	705	c) 696	694
714	712	708	703	698

←20→

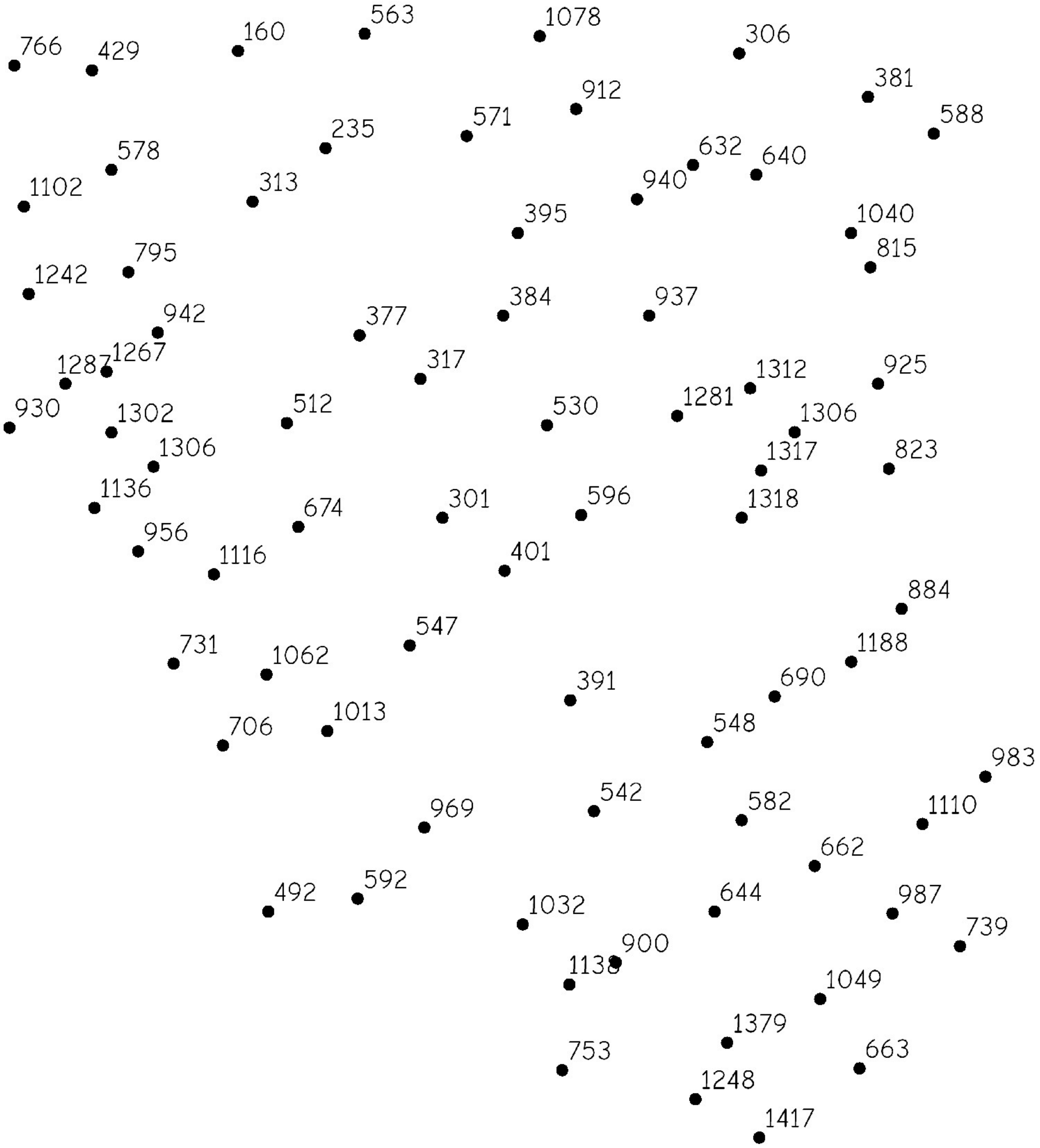
11.10 - Plot a graph of slope in degrees (on x axis) against slope in percent (y axis). Which is usually larger, slope as degrees, or slope expressed as percent?

11.11 - What is an elevation contour?

11.12 - Draw the approximate location of contours for the following set of points. Start contours at the 960 value and use a 30 unit contour interval. For this exercise, it is permissible to estimate the contour locations visually; you do not have to calculate the distances between points to place the contour lines.



11.13 - Draw the approximate location of contours for the following set of points. Start contours at the 0 value and use a 200 unit contour interval. For this exercise, it is permissible to estimate the contour locations visually; you do not have to calculate the distances between points to place the contour lines.



11.14 - What is the formula to calculate a contour height from two measured elevations?

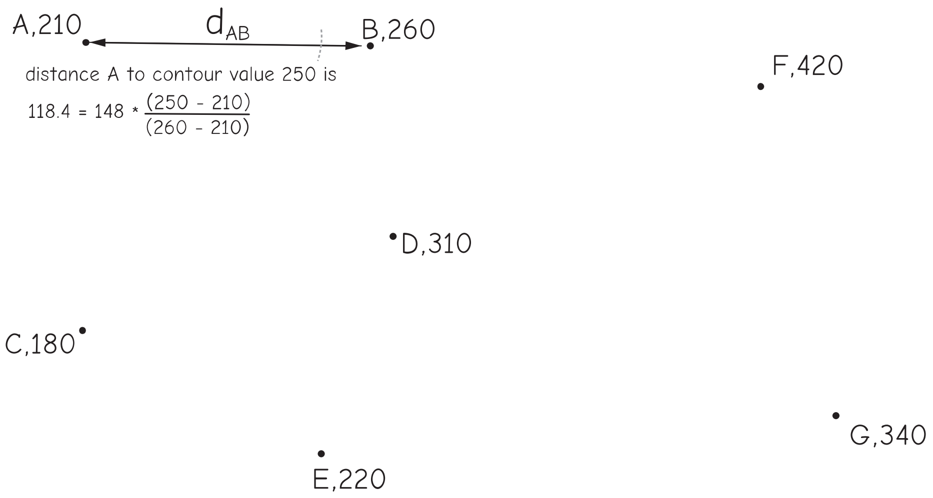
11.15 - Using the figure below, calculate the distances to the listed contour line along the shortest path between points. The example shows the distance calculation from Point A to the contour with value 250, along the straight line from A to B, when the values of A and B are shown, and the distance from A to B is 148.

Distance from B to the 300 contour on the line B - D, when d_{BD} is 94

Distance from E to the 250 contour on the line E - D, when d_{ED} is 115

Distance from C to the 200 contour on the line C - D, when d_{CD} is 188

Distance from E to the 300 contour on the line E - G, when d_{EG} is 248



11.16 - Using the figure above, answer the following:

Distance from A to the 200 contour on the line A - C, when d_{AC} is 94

Distance from E to the 300 contour on the line E - D, when d_{AC} is 115

Distance from F to the 400 contour on the line F - G, when d_{FG} is 178

Distance from B to the 350 contour on the line B - F, when d_{BF} is 224

Distance from E to the 250 contour on the line E - G, when d_{EG} is 248

11.17 - What are the plan curvature and profile curvature, and how do they differ?

11.18 - Define the watershed boundaries and possible stream locations in the digital elevation data depicted below:

373	366	369	383	378	356	337	329	328	326	327	326	331	338	330	322	322	314	301	293
384	380	384	392	380	357	343	339	340	341	342	343	346	350	343	335	327	312	303	304
409	405	405	401	380	360	354	356	361	362	359	356	354	352	349	345	336	320	316	322
420	417	416	407	384	367	368	375	376	369	357	345	337	332	332	335	336	327	321	323
399	397	401	399	379	362	367	381	381	365	344	327	317	312	312	320	328	323	313	310
369	369	377	381	366	349	353	370	378	366	343	324	310	300	301	315	328	320	306	300
355	353	362	370	359	340	338	357	372	365	344	324	305	293	296	315	329	319	298	288
349	343	348	359	351	332	327	342	358	357	341	318	295	285	292	310	322	313	292	278
343	333	336	348	338	318	313	323	341	349	335	309	283	275	282	298	310	301	282	270
336	323	326	336	324	304	297	305	326	339	329	302	275	265	270	285	298	291	271	260
322	309	308	321	316	297	283	289	309	325	322	299	271	256	260	274	286	281	262	249
308	295	292	309	314	295	275	273	288	303	302	286	265	250	250	260	266	263	249	240
298	287	281	295	305	293	273	262	269	277	274	264	253	242	239	242	245	244	238	233
282	275	269	275	283	281	269	257	254	255	251	244	238	234	231	231	232	232	229	228
278	275	270	265	264	264	262	257	252	245	240	238	237	236	235	234	234	233	229	225
303	299	289	274	266	262	257	256	252	246	246	251	253	252	250	249	248	244	236	228
322	313	296	280	278	276	265	257	251	249	261	272	273	273	271	271	267	259	247	234
321	308	290	279	285	287	276	265	253	252	269	282	284	284	284	285	280	268	253	238

11.19 - Define the watershed boundaries and possible stream locations in the digital elevation data depicted below:

162	108	67	103	56	66	130	214	153	122	70	36	56	91	165
169	160	101	120	95	115	119	202	212	121	55	43	101	158	261
254	224	182	158	214	142	208	249	225	129	58	121	137	253	344
323	312	204	191	214	228	300	345	195	126	58	105	188	298	381
338	334	267	307	231	194	200	190	176	114	63	141	199	277	278
438	471	405	344	228	242	194	137	103	81	111	103	198	262	195
550	550	387	304	301	330	245	257	175	110	163	204	225	206	144
669	557	502	414	451	378	396	329	180	148	242	349	293	191	148
604	639	490	442	433	425	406	264	169	169	278	401	297	241	167
742	666	536	443	340	294	265	202	221	227	339	342	260	260	245
799	630	509	438	456	414	304	344	337	322	359	377	387	375	308
767	685	608	578	457	426	318	442	371	421	430	330	275	292	226
734	789	721	578	512	421	443	512	506	503	378	315	227	213	173
668	765	826	728	579	558	489	534	513	366	330	244	266	190	170
705	767	784	785	761	675	607	545	440	275	226	202	165	104	55

11.20 - Define the following: solar zenith angle, solar azimuth angle, and solar incidence angle.

11.21 - Draw a diagram illustrating the solar incidence angle, and identify what site/terrain factors affect the solar incidence angle.

11.22 - What are viewsheds, when are they used, and how are they calculated?

11.23 - What is a shaded relief map? How are the values for each cell of a hillshade surface calculated?

12 Spatial Estimation: Interpolation, Prediction, and Core Area Delineation

Introduction

Spatial prediction methods are used to estimate values at unsampled locations (Figure 12-1). An obvious question is, why estimate? Why not just measure the value at all locations? Predictions are required because time and money are limiting. There is an infinite number of potential sampling locations for any continuous variable in any study area, and it is impossible to measure at all locations. While there is a finite number of discrete objects in all studies, there are usually too many to measure them all. Practical constraints usually limit samples to a subset of the possible lines, polygons, points, or raster cell locations.

Spatial prediction may be required for other reasons. Besides cost, some areas may be difficult or impossible to visit. A parcel owner may prohibit entry. It may be too dangerous to collect samples, for example, in part of a park because lions may eat the sampling crew, or elephants trample them.

Spatial prediction may be required due to missing or otherwise unsuitable samples. If it is difficult, expensive, or the wrong season for sampling, it may be impossible to replace lost samples. Samples may be discovered as unreliable or suspect once the measuring crew has returned. Suspect, "outlier," points are often dropped from data sets. These now missing points may be cru-

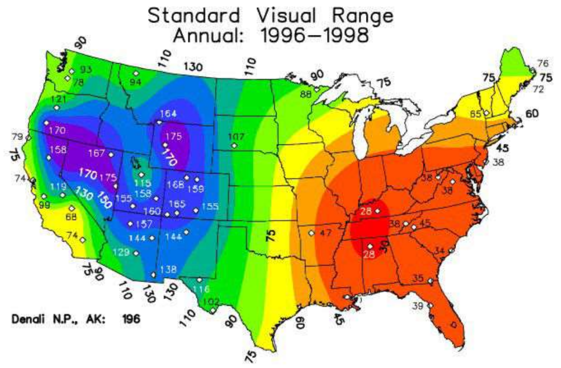


Figure 12-1: Air quality varies across space and time, and is only measured in a few locations. Spatial estimation methods are commonly used to predict air quality at unsampled locations (courtesy U.S. NPS).

cial to the analysis and, if so, the missing values estimated. Finally, estimates may be required when changing to a smaller cell size in a raster data set. The "sampling" frequency is set by the original raster, and values must be estimated for the new, smaller cells.

Spatial interpolation is the prediction of variables at unmeasured locations, and based on a sampling of the same variables at known locations. Most interpolation methods rely on the nearest points to estimate missing values, and use some measure of distance from known to unknown values.We might have measured air pollution at a set of towers across a region, but need estimates for all locations in that region. Interpolation is routinely used to estimate air and water temperature, soil moisture, elevation, ocean productivity, population density, and a host of additional variables.

Spatial prediction also involves the estimation of variables at unsampled locations, but differs from interpolation in that estimates are based at least in part on other variables, and often on a total set of measurements. We may use elevation to help estimate temperature because it is often cooler at higher locations. A map of elevations may be combined with a set of measured temperatures to estimate temperatures at unknown locations.

A *core area* is characterized by high use, density, intensity, or probability of occurrence for a variable or event. Core areas are defined from a set of samples, and are used to predict the frequency or likelihood of occurrence of an object or event. Home ranges for individual animals, concentrations of business activity, or centers of criminal activity are all examples of core areas. There are several methods that may be used in identifying these core areas. These methods typically draw from a set of sample points that constitute events, such as an observation of an animal, a business location, or a crime that has been committed.

Spatial prediction typically translates from lower spatial dimensions to the same or higher dimensions. This means we typically generate points or lines from point data, or areas from point, line, or area data. Prediction methods allow us to extend the information we have collected, most often to "fill in" between sampled locations, but also to improve the quality of the data we have collected.

Spatial prediction methods may also be used to translate information from a higher order to a lower order, that is, to estimate point values from data collected or aggregated to area or lines. We may have population data reported for an area, and we may wish to estimate population for a specific point within this area. This may be affected by the *modifiable areal unit problem*, a common hazard in spatial estimation methods that was described in Chapter 9.

Whatever the methods used, spatial estimation is based on a set of samples. An individual sample consists at least of the coordinates of the sample location and a measurement of the variable of interest at the sample location. We may also measure additional, related variables at the sample location. Coordinates should be measured to the highest accuracy and precision practical, given cost and time constraints and the intended use of the data. Sample variables should be measured using accurate, standardized, repeatable methods.

Sampling

Estimation is based on a sample of known points. The aim is to estimate the values for a variable at unknown locations based on values measured at sampled locations. Planning will improve the quality of the samples, and usually leads to a more efficient and accurate interpolation.

We control two main aspects of the sampling process. First, we may control the location of the samples. Samples must be spread across our working area. However, we may choose among different patterns in dispersing our samples. The pattern we choose will in turn affect the cost of our samples and the quality of our interpolation. A poor distribution of sample points may increase errors or may be inefficient, resulting in unnecessary costs.

Sample number is the second main aspect of the sampling process we control. One might believe the correct number is "as many as you can afford;" however, this is not always the case. A law of diminishing returns may be reached, and further samples may add relatively little information for substantially increased costs. Unfortunately, in most practical applications, the available funds are the main limiting factor. Most surfaces are undersampled, and additional funds and samples would usually increase the quality of the interpolated surface. To date, there have been relatively few studies or well-established guidelines for determining the optimum sample number for most interpolation methods.

There are times when we control neither the distribution nor the number of sampling points. This often occurs when we are working with "found" variables, for example, the distribution of illness in a population. We may identify the households where a family member has contracted a given illness. Although we can control neither the number nor the distribution of ill people, we may wish to use these "samples" in an interpolation procedure.

Sampling Patterns

There are several commonly applied sampling patterns. A *systematic sampling pattern* is the simplest (Figure 12-2a), because samples are spaced uniformly at fixed X and Y intervals. The intervals may not be the same in both directions, and the X and Y axes are not required to align with the northing and easting grid directions. The sampling pattern often appears as points placed systematically along parallel lines.

Systematic sampling has an advantage over other sampling patterns by way of ease in planning and description. Field crews quickly understand how to lay out the sample pattern, and there is little subjective judgement required.

However, systematic sampling may have disadvantages. It is usually not the most statistically efficient sampling pattern because all areas receive the same sampling intensity. If there is more interest or variation in certain portions of the study area, this preference is not addressed by systematic sampling. The difficulty and cost of traveling to the sample points is not considered. It may be difficult or impossible to stay on line between sampling points. Rough terrain, physical barriers, or lack of legal access may preclude sampling at prescribed locations.

In addition, systematic sampling may introduce a bias, particularly if there are patterns in the measured variable that coincide with the sampling interval. For example, there may be a regular succession of ridges and valleys associated with underlying geologic conditions. If the systematic sampling interval coincides with this pattern, there may be a bias in sample values.

Random sampling (Figure 12-2b) may avoid some, but not all, of the problems that affect systematic sampling. Random sampling entails selecting point locations based on random numbers. Typically, both the easting and northing coordinates are chosen by independent random processes. These

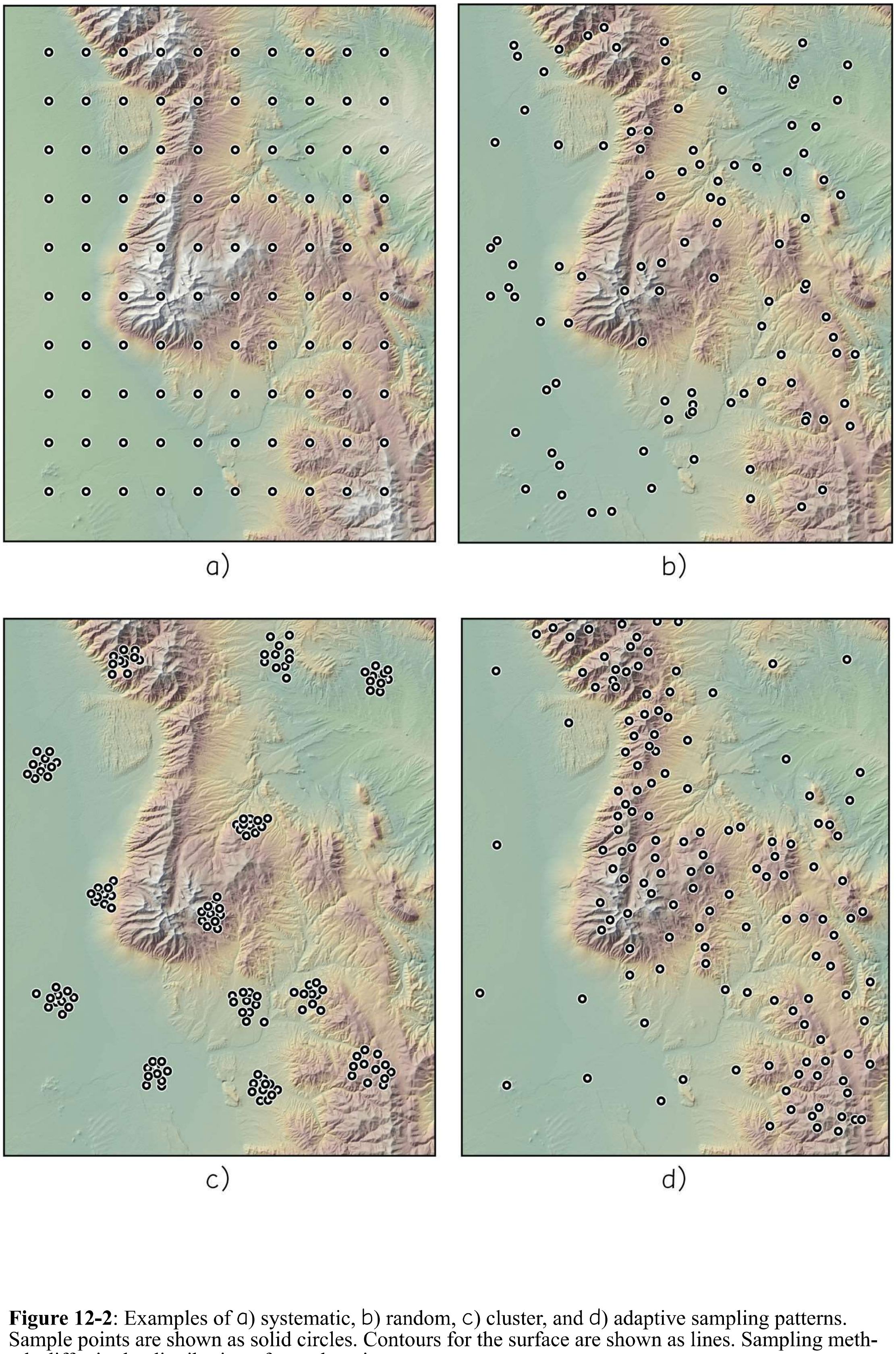


Figure 12-2: Examples of a) systematic, b) random, c) cluster, and d) adaptive sampling patterns. Sample points are shown as solid circles. Contours for the surface are shown as lines. Sampling methods differ in the distribution of sample points.

may be plotted on a map and/or listed, and then visited with the aid of a GNSS or other positioning technology to collect the sample. The points do not have to be visited in the order in which they were selected, so in some instances, travel distances between points will be quite small. On average, the distances will be no shorter than with a systematic sample, so travel costs are likely to be at least no worse than with systematic sampling.

Random samples have an advantage over systematic samples in that they are unlikely to match any pattern in the landscape. Hence, the chances are lower for biased sampling and inaccurate predictions.

However, like systematic sampling, random sampling does nothing to distribute samples in areas of high variation. More samples than necessary may be collected in uniform areas, and fewer samples than needed may be collected in variable areas. In addition, random sampling is more complicated and hence more difficult to understand than systematic sampling. More training may be required for sampling crews when implementing random sampling. Random sampling is seldom chosen when sampling over large areas, due to these disadvantages and relatively few advantages over alternative sampling strategies.

Cluster sampling is a technique that groups samples (Figure 12-2c). Cluster centers are chosen by some random or systematic method, with a cluster of samples arranged around each center. The distances between samples within a cluster are generally much smaller than the distances between cluster centers.

Reduced travel time is the primary advantage of cluster samples. Travel times withing a cluster are shorter. A sampling crew may travel several hours to reach a cluster center, but only a few minutes between each sample within a cluster. Cluster sampling is often used in natural resource surveys that entail significant off-road travel because of the reduction in travel times.

There are several variants of cluster sampling. Cluster centers may be located randomly or systematically. Samples within a cluster may also be placed at random or systematically around the cluster center. Both approaches have merit, although it is more common to locate cluster centers at random and distribute samples within a cluster according to some systematic pattern. This approach is used by the U.S. Forest Service to conduct national surveys of forest conditions, and by many prospectors during mineral exploration.

Adaptive sampling is a final method we will describe. Adaptive sampling is characterized by frequent sampling in variable areas and sparse sampling in uniform areas (Figure 12-2d). Adaptive sampling greatly increases sampling efficiency because small-scale variation is better sampled. Large, relatively homogeneous areas are well represented by a few samples, reserving more samples for areas with higher spatial variation.

Adaptive sampling requires a way to estimate feature variation prior to field visits or while in the field, or repeat visits to the sampling areas. Sample density is adaptively increased in areas of high variation. Some times it is quite obvious where the variation is greatest while in the field. For example, when measuring elevation, it is obvious where the terrain is more variable. Sample density may be increased based on field observations of steepness.

If there is no method of identifying where the features are most variable while in the field, then sample density cannot be increased "on the spot." Samples may require office or lab for analysis to estimate variation. Sample locations are then selected based on local variation. The list or map of coordinate locations may be generated and used as a guide in collecting subsequent samples.

Spatial Interpolation Methods

There are many different interpolation methods. While methods vary, all combine the sampled values and positions to estimate values at unmeasured locations. Mathematical functions are used that incorporate distance between the interpolation points and the sample points with the values at the sample points. Methods differ in the mathematical functions used to weight each observation, and the number of observations used. Some interpolators use every observation when estimating values at unsampled locations, while other interpolators use a subset of samples, for example, the three points nearest an unmeasured location.

Different interpolation methods will often produce different results, even when using the same input data. This is due to the differences in the mathematical functions and number of data points used when estimating values for the unsampled locations. Each method may have unique characteristics, and the overall accuracy of an interpolation will often depend on the method and samples used.

Accuracy is often judged by the difference between the measured and interpolated values at a number of withheld sample points. These withheld points are not used when performing the interpolation, but are checked against the interpolated surface. However, no single interpolation method has been shown to be more accurate than all others for every application. Each individual or organization should test several sampling regimes and interpolation methods before adopting an interpolation method.

Interpolation methods may produce one or more of a number of different output types. Interpolation is usually used to estimate values for a raster data layer. Other methods produce contour lines. Contour lines are less frequently produced by interpolation methods, but are a common way of depicting a continuous surface. At least one interpolation method defines polygon boundaries.

Interpolation to a raster surface requires estimates of unmeasured values at the center of each raster grid cell. Raster layer boundaries and cell dimensions are specified, in turn defining the location of each raster cell.

We will describe the most common interpolation methods and apply them all to a single data set to facilitate comparisons. Figure 12-3 shows sample points for ozone data for the eastern United States, collected by various health and environmental agencies, and an index value for the 2014 year. Denver is to the extreme left, New England in the upper right, and Atlanta indicated by the cluster of sampling near the lower right of this figure. Circles are sized and colored to reflect the 98th percentile measurement, in parts per billion (ppb) during daytime hours, a value related to injury caused by ozone exposure. Weather, combustion, chemical release, and topographic conditions can combine to create hazardous concentrations, particularly for vulnerable populations. Since it is expensive and difficult to make precise ozone measurements, the network is limited, and there is a need to interpolate between sampling stations. These sample points will be used to demonstrate the application of various interpolation and spatial prediction methods in the following sections of this chapter. Estimated ozone concentration surfaces for each method will be shown.

Note that the comparisons and figures are only to illustrate different interpolation methods. They are not to establish the relative merit or accuracy of the various methods. The best interpolation method for any given application depends on the characteristics of the variable to be estimated, the cost of sampling, available resources, and the accuracy requirements of the users.

An independent error measure is required if we are to obtain a good estimate of the interpolation accuracy. Accuracy estimates may be obtained with a withheld sample technique, where the surface is fit to the

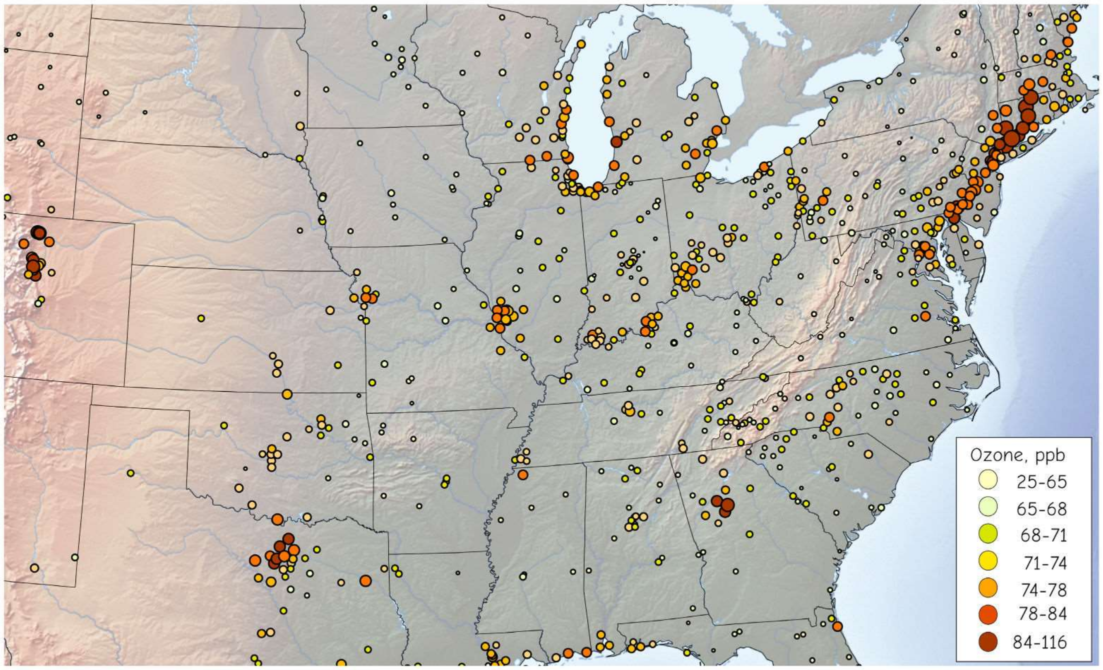


Figure 12-3: Measurement points and values for indicator ozone concentration for 2014. Higher concentrations harm public health, with readings over 85 parts per billion associated with damage to lung function, particularly in the elderly, children, and asthmatics. These sample data will be used later in this chapter to demonstrate interpolation methods.

data withholding one data point. The error is estimated at the withheld point as the observed minus the interpolated values. The sample is replaced, a new sample selected and withheld, and the surface fit and error again determined. This is repeated for each data point. A less efficient testing method entails collecting an independent set of sample points that are withheld from the interpolation process. Their measured values are then compared to the interpolated values, and the mean error, maximum error, and perhaps other error statistics identified.

Nearest Neighbor Interpolation

Nearest neighbor interpolation, also known as Thiessen polygon interpolation, assigns a value for any unsampled location that is equal to the value found at the nearest sample location. This is conceptually the simplest interpolation method, in the sense that the mathematical function used is the equality function, and only one point, the nearest point, is used to assign a value to an unknown location.

The nearest neighbor interpolator defines a set of polygons, known as Thiessen polygons. All locations within a given Thiessen polygon have an identical value for the Z variable (in this and other chapters, Z will be used to denote the value of a variable of interest at an X and Y sample location). Z may be elevation, size, production, or any other variable we may measure at a point. Thiessen polygons define a region around each sampled point that have a value equal to the value at the nearest sampled point. The transition between polygon edges is abrupt; that is, the variable jumps from one value to the next across the Thiessen polygon boundary.

The three-dimensional perspective representation of an interpolated nearest neighbor surface illustrates some characteristics of output surfaces (Figure 12-4). Heights in the figure correspond to the input values at the points. The polygon has a uniform value that corresponds to the input sample value. Polygons are of irregular size, and values change abruptly along the polygon edges.

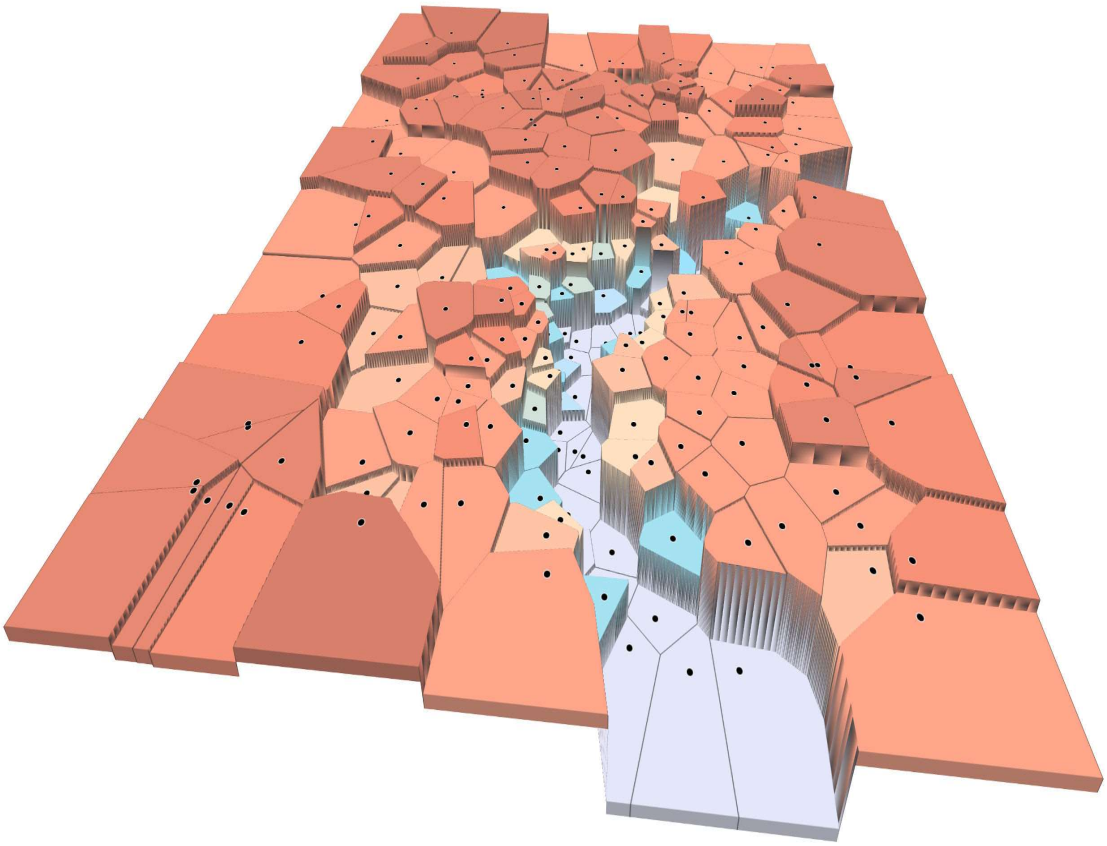


Figure 12-4: A three-dimensional perspective representation of Thiessen (nearest neighbor) polygons, generated by interpolation for a set of sample points (black dots). Note that areas are assigned the value of the nearest sample point, creating a set of irregular polygons. Values change abruptly along the polygon edges.

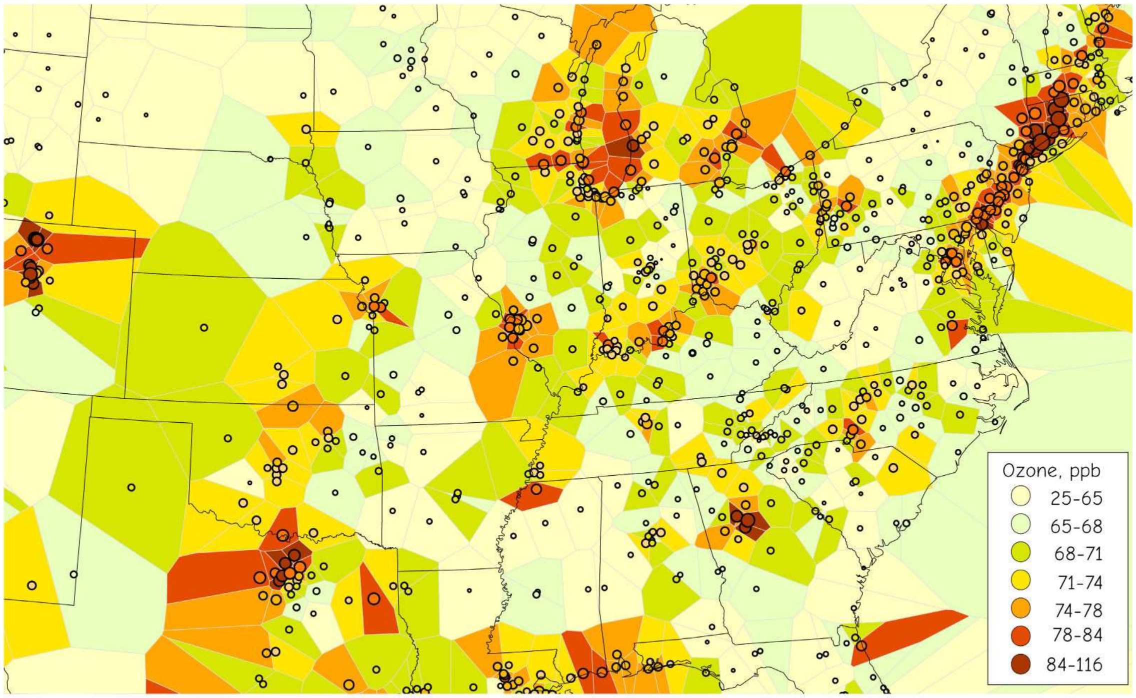


Figure 12-5: Sample points and estimated ozone concentrations by Thiessen polygons (right).

Figure 12-5 shows our ozone sample points and Thiessen polygons based on the sample points. Note that sampling is denser near some urban areas, particularly the Philadelphia–New York City corridor in the upper right, Denver on the left, and St. Louis, Dallas, and Atlanta along the mid to lower portions of the figure. Thiessen polygons are smaller where sampling density is highest.

Thiessen polygons provide an *exact interpolator*. This means the interpolated surface equals the sampled values at each sample point. The value for each sample location is preserved, so there is no difference between the true and interpolated values at the sample points. Exact interpolators have this admirable quality, but often are not the best interpolators at unsampled points; for example, the Thiessen polygon method is usually in error at nonsampled locations, often more so than other inexact interpolators.

Fixed Radius – Local Averaging

Fixed radius interpolation is more complex than nearest neighbor interpolation, but less complex than most other interpolation methods. In a fixed radius interpolation, a raster grid is specified in a region of interest. Cell values are estimated based on the average of nearby samples.

The samples used to calculate a cell value depend on a *search radius*. The search radius defines the size of a circle that is centered on each cell. Sample points found inside the circle are averaged to interpolate the value for that cell (Figure 12-6). Points outside the circle are ignored.

Figure 12-7 shows a perspective view of fixed radius sampling. Note that there is a sample data layer, shown at the top of Figure 12-7, vertically aligned with the interpolated surface. This surface is a raster data layer with interpolated values in each raster cell. A fixed radius circle is centered over a raster cell. The average is calculated for all samples contained within the sample circle, and this average is placed in the appropriate out-

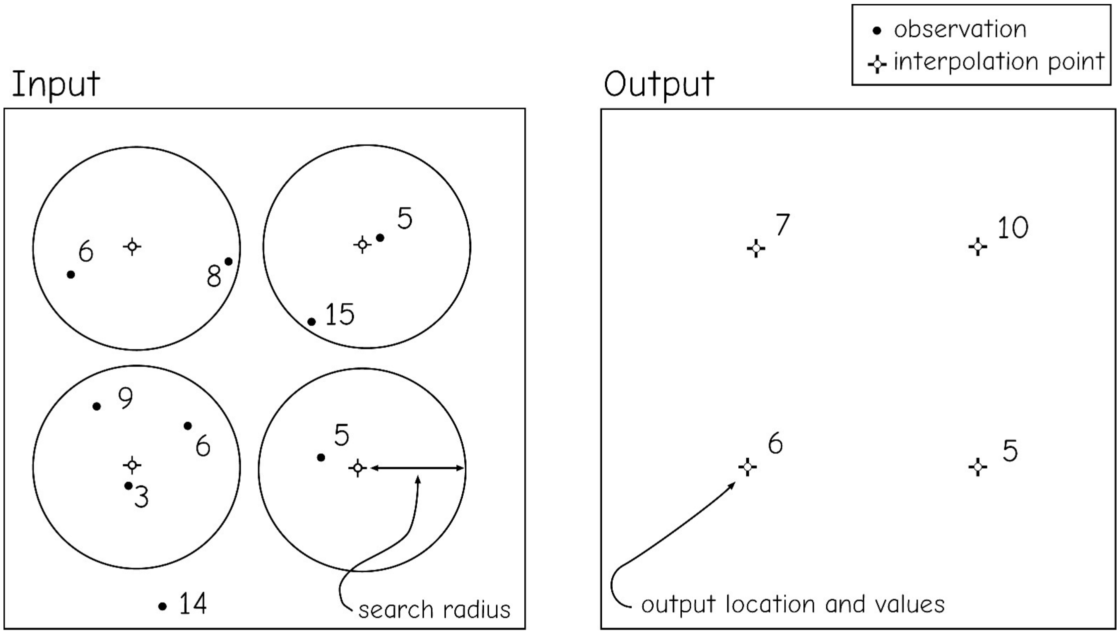


Figure 12-6: A diagram and example of a fixed radius interpolation. Values within each sampling circle are averaged to estimate an output value for the corresponding point.

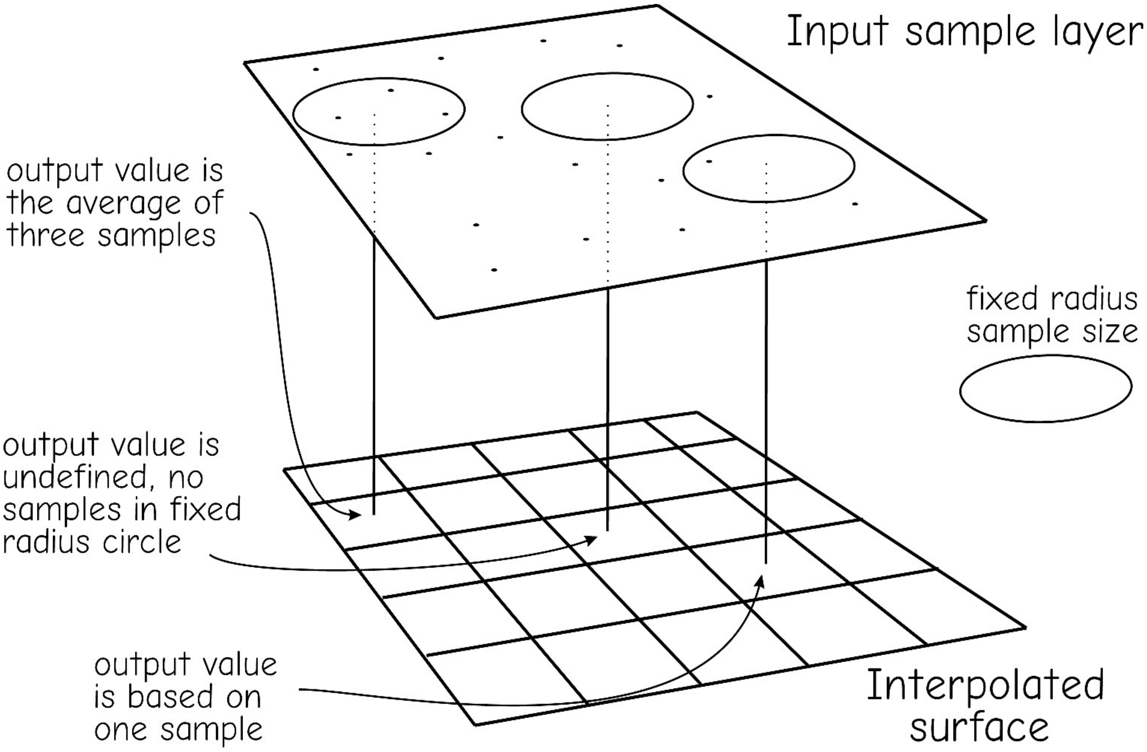


Figure 12-7: A perspective diagram of fixed radius sampling. A circle is centered on each raster cell location. Samples within the circle contribute to the value assigned to each corresponding raster cell (adapted from Mitchell, 1999).

put raster cell. The process is repeated for each raster cell in the surface. The fixed radius circles are shown corresponding to three raster cells, containing three, zero, and one sample points, respectively. Circles may contain no points, in which case a zero or no data value is placed in the raster cell. The radius for the circle is typically much larger than the raster cell's width. This means circles overlap for adjoining cells, causing neighboring cell values to be similar.

The fixed radius interpolator tends to smooth the sample data (Figure 12-8). Large or small values sampled at a given point are maintained when only that one sample point falls within a search radius for a cell. Values are brought toward the overall sample mean when averaged within a search radius.

The search radius affects the values of the interpolated surface. Too small a search radius results in many empty cells, with no data or null values. Too large a search radius may smooth the data too much. In the extreme case, a search radius may be defined that includes all sample points for all cells, resulting in a single interpolated value for all cells. Some intermediate search radius is chosen.

Fixed radius interpolators are not exact interpolators because they may average several points in the vicinity of a sample, and so they are unlikely to place the measured value at sample points in the interpolated surface.

Inverse Distance Weighted Interpolation

The inverse distance weighted (IDW) interpolator estimates the value at unknown points using the sampled values and distance to nearby known points. The weight of each sample point is an inverse proportion to the distance, thus the name. The farther away the point, the less weight the point has in

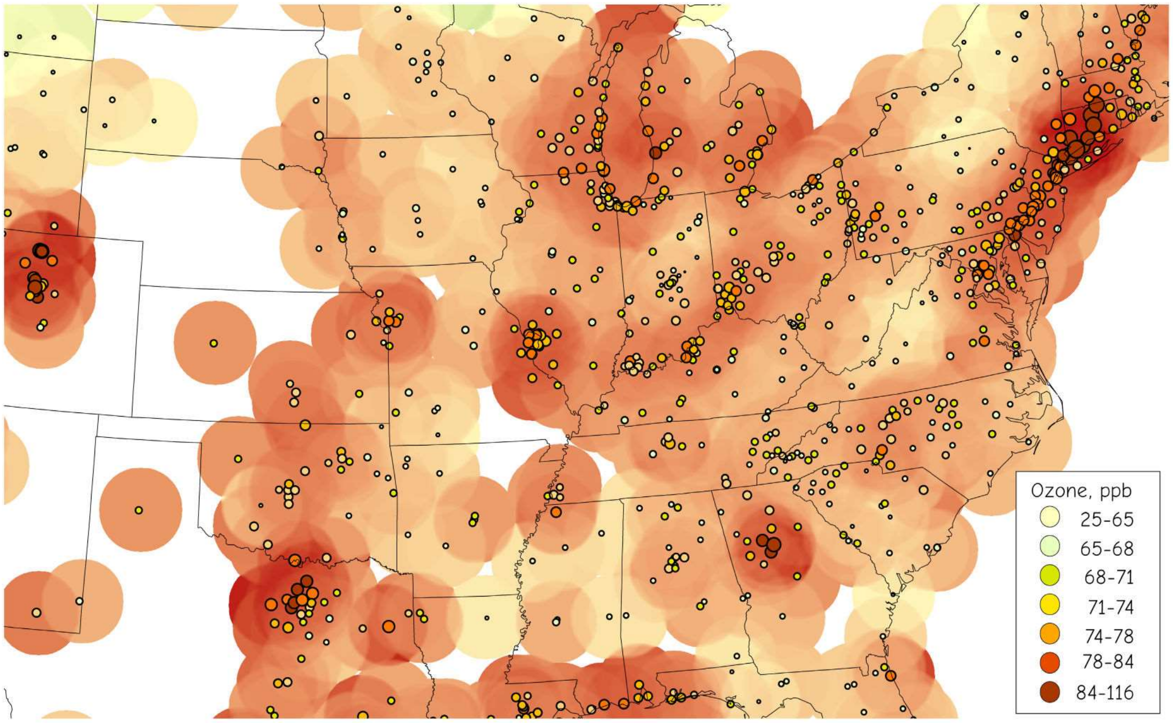


Figure 12-8: Original data and sample points, and a fixed radius interpolation. Note that this method may leave gaps in areas of sparse samples, and average over dense sample areas. Increasing the averaging radius will decrease gaps, but increase data smoothing in highly sampled areas.

helping define the value at an unsampled location. Values are estimated by:

$$Z_j = \frac{\sum_i \frac{Z_i}{d^n_{ij}}}{\sum_i \frac{1}{d^n_{ij}}} \qquad (12.1)$$

where Z_j is the estimated value for the unknown point at location j, d_{ij} is the distance from known point i to unknown point j, Z_i is the value for the known point i, and n is a user-defined exponent. Any number of points greater than two may be used, up to all points in the sample. Typically, some fixed number of close points is used; for example, the three nearest sampled points will be used to estimate values at unknown locations. Note that n controls how fast a point's influence wanes with distance. The larger the n, the smaller the weight ($1/dn_{ij}$), so the less influence a point has on the estimate of the unknown point.

Figure 12-9 illustrates an IDW interpolation calculation. The three nearest samples are used. Each measured sample value is

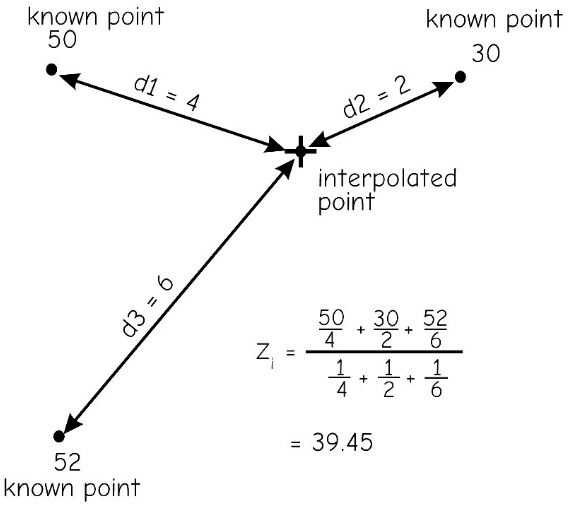


Figure 12-9: An example calculation for a linear inverse distance weighted interpolator. The values at each known point (50, 52, 30) are averaged, with weights based on the distances (d1, d2, d3) from the interpolated point.

weighted by the inverse of the distance from the unknown, interpolated location. These weighted values are added. The result is divided by the sum of the weights to "scale" the weights to the measurement units. This produces an estimate for the unsampled location.

IDW is an exact interpolator. Interpolated values are equal to the sampled values at each sampled point. As a d_{ij} becomes very small (sample points near the interpolated location), the $1/d_{ij}$ becomes very large. The contribution from the nearby sample point dwarfs the contributions from all other points. The values $1/dij$ are very near zero for all i values except the one very near the sampled point, so the values at all other points are effectively multiplied by zero in the numerator of the IDW equation. The sum in the denominator reduces to the weight $1/dij$. The weights on the top and the bottom of the IDW equation become more similar, and the fraction approaches 1. Thus, at a sampled point the IDW interpolation formula reduces to:

$$\frac{\frac{Z_i}{d^n_{ij}}}{\frac{1}{d^n_{ij}}} \qquad (12.2)$$

By simple division this is reduced mathematically to Z_i, the value measured at the sampling location.

Inverse distance weighting results in smooth interpolated surfaces (Figure 12-10). The values do not jump discontinuously at edges, as occurs with Thiessen polygons, and sometimes with fixed radius interpolation. While IDW is easily and widely applied, care must be taken in evaluating the particular n and i selected. The effects of changing n and i should be tested in an over-sampled case or using retention and repeat fitting methods, described later, where adequate withheld points can be compared to interpolated points. The IDW, and all other interpolators, should be applied only after the user is convinced the method provides

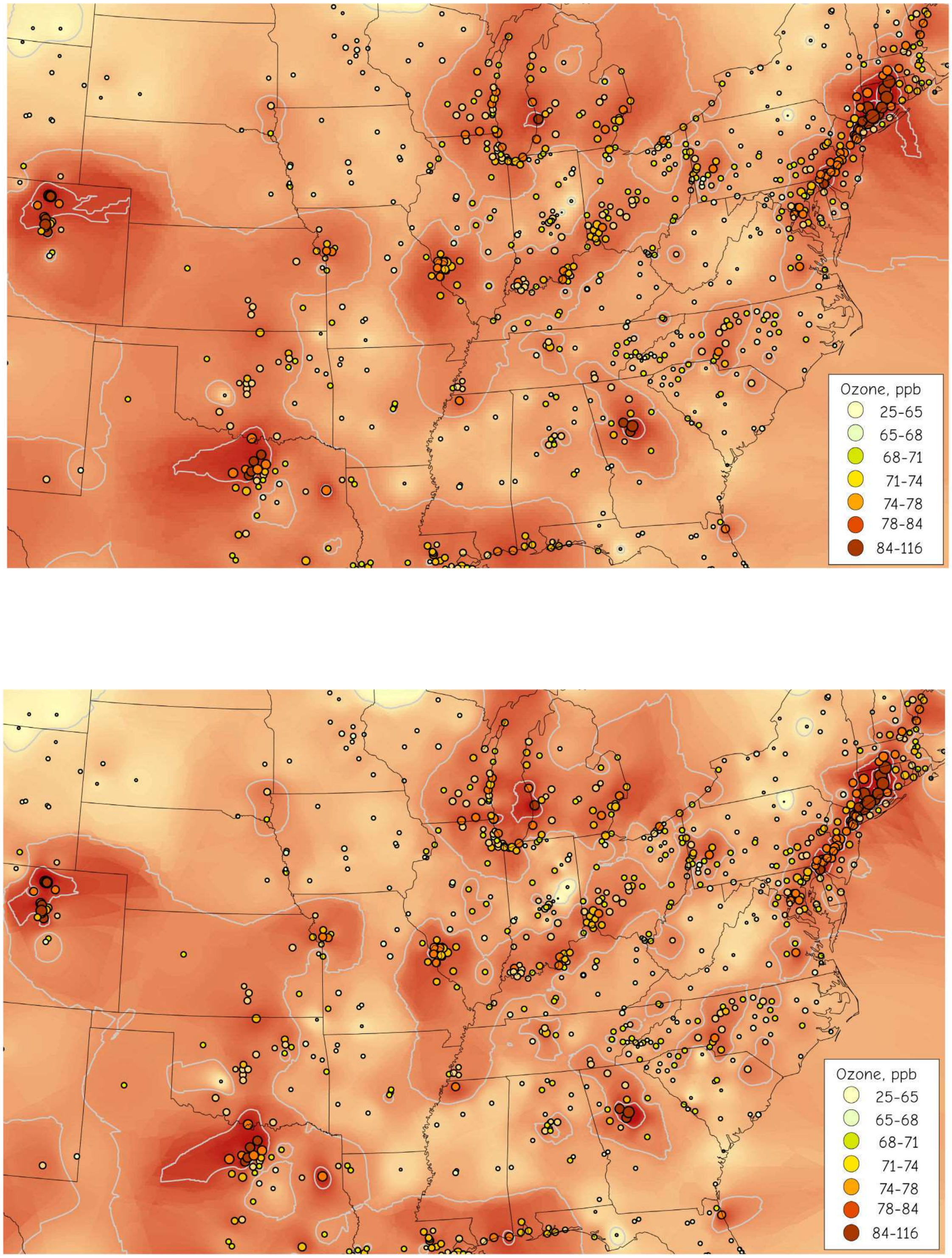


Figure 12-10: Inverse distance weighted interpolation typcially results in a smooth, gap-free surface. Exponent order and sample size affect the surface shape for this interpolator. The top interpolation relies on the 12 nearest points, with an exponent of two, while the lower interpolation relies on the four nearest points, and an exponent of three. Local influences are stronger as the exponent increases and the number of sample points decreases.

estimates with sufficient accuracy. In the case of IDW, this may mean testing the interpolator over a range of n and i values, and selecting the combination that most often gives acceptable results.

The size of the user-defined exponent, n, affects the shape of the interpolated surface (Figure 12-10). When a larger n is specified, the closer points become more influential. Higher exponents result in surfaces with higher peaks, lower valleys, and steeper gradients near the sampled points. Contours become much more concentrated near sample points when $n = 2$ (Figure 12-10, top) than when $n = 3$ (Figure 12-10, bottom). These changing shades reflect steeper gradients near the known data points.

The number of points, i, used to estimate an interpolated point, j, also affects the estimated surface, but effects are often complex and difficult to generalize, because they depend on the distribution and magnitudes of the specific sample points. A larger number of sample points tends to result in a smoother interpolated surface.

Splines

A *spline* is a flexible ruler that was commonly used by draftsmen to create smooth curves through a set of points. Mathematical spline functions, also referred to as splines, are used to interpolate along a smooth curve. These functions serve the same purpose as the flexible ruler in that they force a smooth line to pass through a desired set of points. Spline functions are more flexible because they may be used for lines or surfaces and they may be estimated and changed rapidly. The sample points are "guides" through which the spline passes.

Spline functions are constructed from a set of joined polynomial functions. Line functions will be described here, but the principles also apply to surface splines. Polynomial functions are fit to short segments. An exact or a least squares method may be used to fit the lines through the points found in the segment. For example, a third-order polynomial may be fit to a line segment (Figure 12-11). A different third-order polynomial will be fit to the next line segment. These polynomials are by their nature smooth curves within a given segment.

Splines are typically first, second, or third order, corresponding to the maximum exponent in the equation used to fit each segment (e.g., second order for x2, third order for x3 or x2y). Segments meet at *knots*, or *join points*. These join points may fall on a sampled point, or they may fall between sampled points.

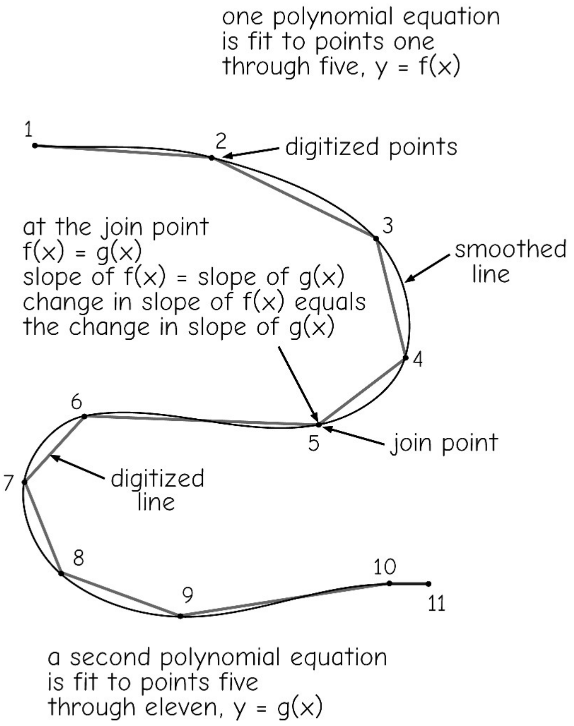


Figure 12-11: Diagram of a two-dimensional (line) spline. Segments are fit to portions of a line. Segments are constrained to join smoothly at knots, where they meet.

Constraints are set on spline functions to ensure the entire line or surface remains smooth at the join points. These constraints are incorporated into the mathematical form of the function for each segment. These constraints require that the slope of the lines and the change in slope of the lines be equal across segments on either side of the join point. Typically, spline functions give exact interpolation (the splines pass through the sample points) and show a smooth transition (Figure 12-12). Strictly enforcing exact interpolation can sometimes lead to artifacts at the knots or between points. Large loops or deviations may occur. The spline functions are often modified to allow some error in the fit, particularly when fitting surfaces rather than lines. This usually removes the artifacts of spline fits, while maintaining the smooth and continuous interpolated lines or surfaces.

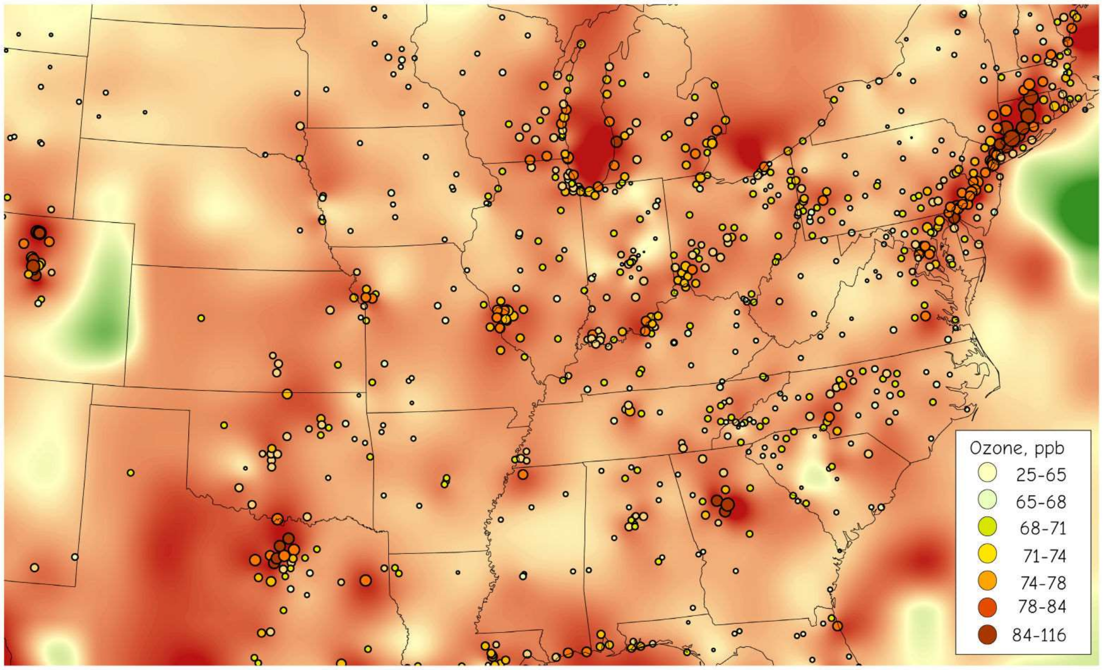


Figure 12-12: A spline-fit surface.

Spatial Prediction

Spatial predictions are based on mathematical models, often built via a statistical process. These statistically based models use coordinate location and measured or observed independent variables to predict values for important but unknown dependent variables. Spatial prediction is different from interpolation because it uses a statistical fitting process rather than a predefined algorithm, and because spatial prediction uses independent variables as well as coordinate locations to estimate unknown variables. We admit that our distinction between spatial prediction and interpolation is artificial, but it is useful in organizing our discussion, and highlights an important distinction between our data-driven models and our fixed interpolation methods.

Spatial predictions are a special case of general predictive modeling, the focus of applied statistics. There is a rich literature devoted to spatial statistics in general, and spatial predictive modeling in particular. We will only scratch the surface of this field; the reader is referred to the introductory spatial statistics texts listed at the end of this chapter.

Our discussions will be restricted to predicting continuous spatial variables. These variables are conceptualized as *spatial fields* that occur across an area, are measured on an interval/ratio scale, and typically have values that vary in concert — that is, they are spatially correlated. This is in contrast to discrete objects, such as point, line, or polygon features. While the occurrence and properties of discrete features may be predicted using spatial models, this is less common, and most discrete object predictions use a different set of tools that will not be discussed here.

Spatial prediction may be considered more general than interpolation. Both are used to estimate values of a target variable at unknown locations. Interpolation methods use only the measured target variable and sample coordinates to estimate the target variables at unknown locations, while spatial prediction usually incorporates additional variables.

Spatial predictions are often improved due to *spatial autocorrelation*. Spatial autocorrelation is the tendency of nearby objects to vary in concert. High values occur together, as do low values. Explanations of this common condition often refer to the observation of Waldo Tobler, that “everything in the universe is related to everything else, but closer things are more related." However, the nature of the correlation may change from one variable to the next, or it may change in space. Correlations may be strong in one region but poor in another, or positive in one area and negative in another. We may improve our predictions if we study the spatial autocorrelation and incorporate the correlation structure into our models.

In addition to spatial autocorrelation, there may be *cross-correlation* between different variables: the tendency for two variables to change in concert. This means two different variables at the same or nearby locations may be high or low together (positive cross-correlation), or highs in one variable correspond to lows in another (negative cross-correlation). Spatial prediction methods may incorporate auto- and cross-correlation in predictions

Surfaces with low and high spatial autocorrelation and with strong cross-correlation are shown in Figure 12-13.. Figure 12-13a shows two surfaces, Layer 1, with a high autocorrelation, and Layer 2, with a low autocorrelation. Scatter diagrams of sample pairs separated by a uniform, short lag distance are shown to the right of each corresponding layer. Higher autocorrelation, as shown in Layer 1, indicates that points near each other are alike. A sample from a surface with high autocorrelation provides substantial information about the values at nearby locations (Figure 12-13a, top). Samples from a surface with low autocorrelation

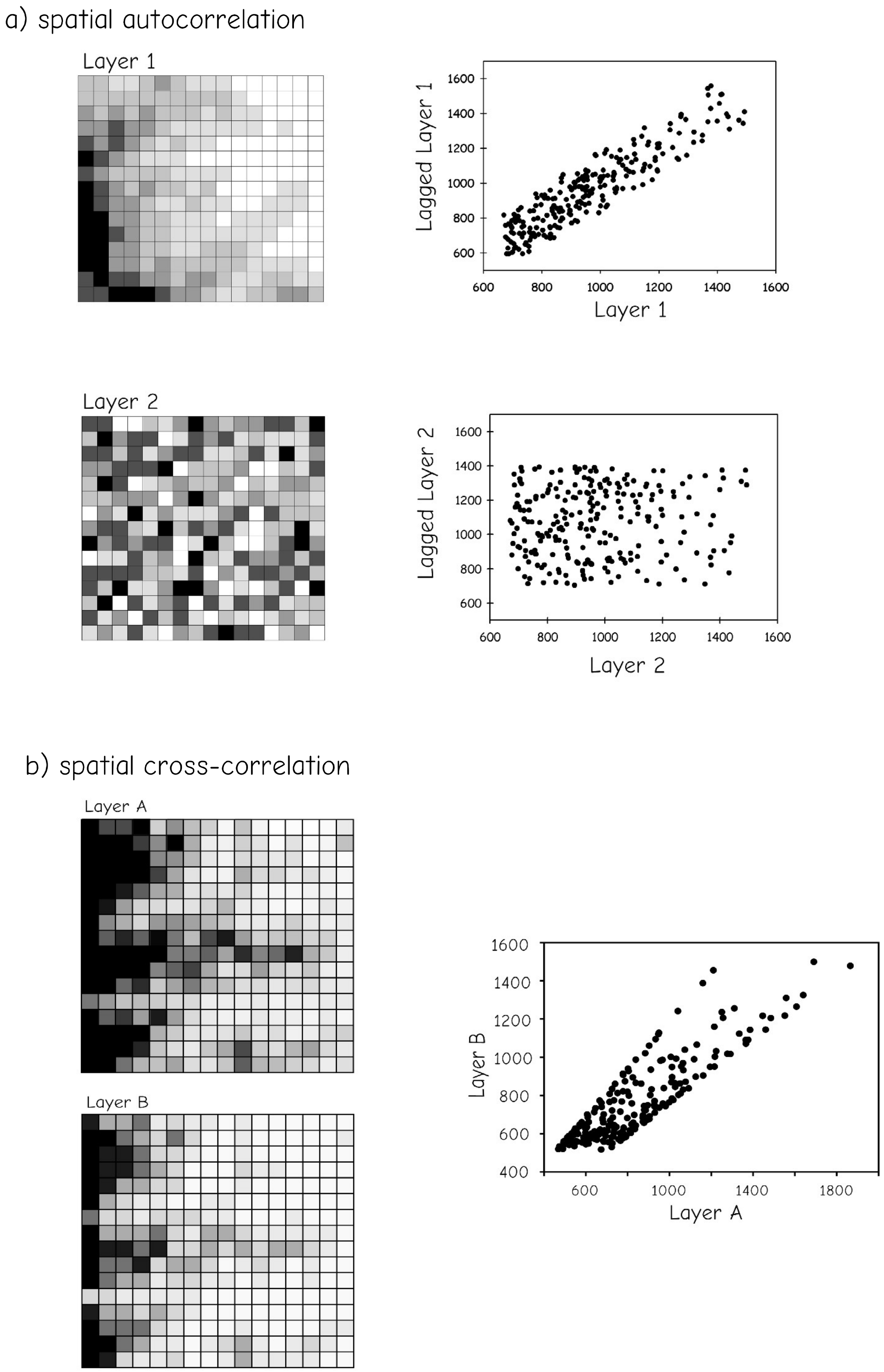


Figure 12-13: Part a shows spatially autocorrelated (Layer 1) and spatially uncorrelated (Layer 2) data layers. Plots of sample pairs with a lag distance h = 1 show similar values for the autocorrelated Layer 1, and unrelated values for uncorrelated Layer 2. Panel b shows two cross-correlated layers. Layer A has higher values on average than Layer B, but the two layers vary in concert. Both reach high and low values in similar areas.

do not provide much information at values in the vicinity of the sample point (Figure 12-13a, bottom).

Two cross-correlated raster layers are shown in Figure 12-13b. Positive cross-correlated layers have values that tend to both be high in some regions and be low in other regions. Many features are positively correlated, such as, housing prices and average income, or donut shop density and number of security guards. Negative cross-correlation occurs when variables change in the opposite sense — areas with high values for one variable are low for the other, for example, low temperatures at higher elevations.

The Moran's I statistic is an established measure of spatial correlation:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i=1}^{n} (Z_i - \bar{Z})^2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \qquad (12.3)$$

where Z_i and Z_j are the variable values at points i and j, respectively; Z is the variable mean; and w_{ij} are weight values that take the value 1 if Z_i and Z_j are adjacent, and 0 if the values are not. An example of Moran's I calculations is shown in Figure 12-14.

Moran's I values approach a value of +1 in areas of positive spatial correlation, meaning large values tend to be clumped together, and small values clumped together. Values

Moran's I is defined as:

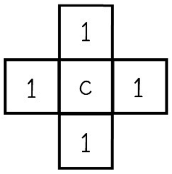
$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i=1}^{n} (Z_i - \bar{Z})^2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}$$

If we impose the rook's case adjacency, then the w_{ij} weights may be represented by the cross below -- 1 on a shared edge for the center cell labeled c, and 0 everywhere else

3	5	2	7	1
9	11	7	3	7
8	(10)	8	3	1
7	12	9	2	7
5	8	1	5	2

The mean value is

$$\bar{Z} = \frac{3+5+2+....1+5+2}{25} = 5.72$$



The calculation of Moran's I for the circled cell is then:

$$= \frac{4[1\cdot(10 - 5.72)(11-5.72)+1\cdot(10 - 5.72)(8-5.72)+1\cdot(10 - 5.72)(12-5.72)+1\cdot(10 - 5.72)(8-5.72)]}{(10 - 5.72)^2\cdot 4}$$

$$= \frac{68.9}{73.3} = 0.94$$

Figure 12-14: Moran's I is a measure of the correlation among nearby observations. This example shows the calculation of Moran's I for a cell in a raster data set. The formula is a weighted combination of the value at a location and neighboring locations. Positive Moran's I values indicate positive spatial correlation, and negative values indicate anticorrelation.

near zero occur in areas of low spatial correlation, and indicate knowing a value at a location does not provide much information about values in adjacent locations — they are just as likely to be different or similar to the observed value. Moran's I approaches -1 when values are anticorrelated — a large value is more likely to be next to small values than next to other large values.

Moran's I may be calculated for both raster and vector data sets. Moran's I values for raster data sets are usually based on the cells immediately adjacent to the focal (center) cell. Weights are typically 1 for cells that share an edge with the focal cell, and 0 otherwise (Figure 12-14). The neighborhood may be "rook's case" and include only cells that share a full edge, as in Figure 12-14, or they may be "king's case" and include all of the eight neighbors in the calculation.

There are many other local indices of spatial autocorrelation, or LISA, including Geary's C, or the Gi of Getis and Ord (1992), and they perform in a manner similar to Moran's I (Anselin, 1995). The indices vary slightly in how they estimate the correlation and in the specific calculations of relatedness and separation. These and a number of additional topics are quite well covered by Anselin (1995), Fotheringham et al. (2000), and O'Sullivan and Unwin (2010), listed in suggested reading at the end of this chapter.

Spatial Regression

Spatial regression and other statistically based models typically use observations of dependent variables, other independent variables, and sample coordinates to develop prediction equations. For example, we estimate temperature across a region using a network of temperature stations. We may interpolate as described in the previous section to estimate temperature, using only the station coordinates and the corresponding temperature measurements. However, we may note a strong cooling trend with elevation, and combine temperature measurements with elevation, latitude, and longitude in a statistical model that provides better temperature predictions. We would then use this model to estimate raster temperature layers for the region.

Spatial predictions are often described mathematically by a general function, such as:

$$Z_i = f(x_i, y_i, \alpha_i, \beta_j) \qquad (12.4)$$

where Z_i is the estimated output value, at the coordinates x_i; y_i at point i; α_i are variables measured at point i; and β_j are variables measured at other locations.

Trend Surface and Simple Spatial Regression

Trend surface prediction is a type of spatial regression that involves fitting a statistical model, or trend surface, through the measured points. The surface is typically a polynomial in the x and y coordinate system. For example, a second-order polynomial model would be:

$$Z = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 y^2 + a_5 xy \qquad (12.5)$$

where Z is the value at any point x and y, and each a_p is a coefficient estimated in a regression model. Least squares methods, described in most introductory statistical textbooks, are used to estimate the best set of a_p values. The a_p values are chosen to minimize the average difference between the measured Z values and the prediction surface.

There must be at least one more sample point than the number of estimated a_p coefficients due to statistical constraints. This does not pose a practical problem for most applications, because the best polynomial models are often second or third order and have fewer than 10 coefficients. More than 10 sample points are typically collected to ensure adequate coverage of a study region.

Trend surfaces are not exact predictors in that the surface typically does not pass through the measured points. There is an error at each sample location, measured as the difference between the interpolated surface and the measurement. Trend surfaces are often among the most accurate methods when fitting smoothly varying surfaces, such as mean daily temperature, over large areas. Trend surfaces typically do not have the "bull's-eye" artifact due to excessive local influence in inverse distance weighted interpolators.

Trend surface methods often perform poorly when there is a highly convoluted surface (Figure 12-15). Ozone as shown in the raw observations can change rapidly over short distances, as can precipitation from a single summer thunderstorm, or population density in a mixed-use neighborhood; this type of abrupt variation is often poorly estimated with a trend surface. Even high-order polynomials may not be sufficiently flexible to fit these complex, convoluted surfaces.

Trend surfaces may be extended to include independent variables that provide some help in predicting the variable of interest:

$$Z = a_0 + a_1x + a_2y + a_3Q + a_4W \quad (12.6)$$

where x and y are the coordinate locations, and Q and W are independent variables measured at the point (x, y), and Z is the depen-

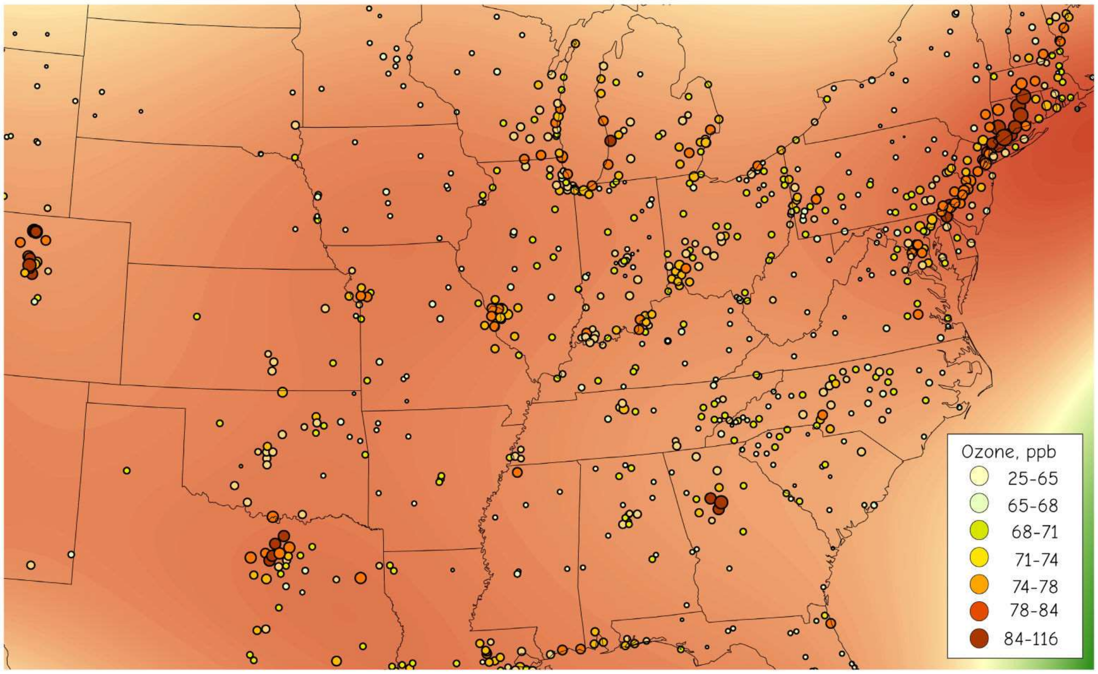


Figure 12-15: A third-order trend surface fit to the sample points.

dent variable to be predicted at the point (X,Y). The a_p values are coefficients for the predictive equation, usually estimated through a least squares statistical process. The value Z may be predicted at any location where we have values for X, Y, Q, and W.

Kriging and Co-Kriging

Kriging is a statistically based estimator of spatial variables. It differs from the trend surface approach in that predictions are based on regionalized variable theory, which includes three main components. The first component is the spatial trend, an increase or decrease in a variable that depends on direction; for example, precipitation may decrease towards the west.

The second component describes the local spatial autocorrelation, that is, the tendency for points near each other to have similar values. Kriging is unique and powerful because we use the observed change in spatial autocorrelation with distance to estimate values at our unknown locations.

The third component in the prediction is random, stochastic variation. These three components are combined in a mathematical model to develop an estimation function. The function is then applied to the measured data to estimate values across the study area.

Much like IDW interpolators, weights in kriging are used with measured sample variables to estimate values at unknown locations. With kriging, the weights are chosen in a statistically optimal fashion, given a specific kriging model and assumptions about the trend, autocorrelation, and stochastic variation in the predicted variable.

Kriging methods are the centerpiece of geostatistics, initially developed in the early 1900s for use in mining. Ore samples may be expensive to obtain or process, and accurate occurrence and density predictions quite difficult and valuable. Kriging estimators were developed to incorporate trends, autocorrelation, and stochastic variation and also provide some estimate of the local variance in the predicted variable.

Kriging uses the concept of a *lag distance,* often symbolized by the letter h. Consider the sample set shown in Figure 12-16. Each value for the variable Z is shown plotted over a region. Individual points may be listed as Z_1, Z_2, Z_3, etc., to Z_k, when there are k sample points. The lag distance for a pair of points is the distance between them, and by convention is denoted by h. The lag distance is calculated from the X and Y coordinate values for the sample points, based on the Pythagorean formula. In our example in Figure 12-16, the lag (horizontal) distance between the locations of sample points Z_1 and Z_2 is approximately 6 units. The difference in values measured at those points, $Z_1 - Z_2$, is equal to 11. Each pair of sample points is separated by a distance, and also has a difference in the values measured at the points. For example, Z_1 is 2.4 units from Z_4, and Z_1 is 5 units from Z_3. Each pair has a given difference in the Z values; for example, Z_1 minus Z_4 is 4. Every possible set of pairs Z_a, Z_b, defines a distance h_{ab}, and is different by the amount $Z_a - Z_b$. The distance h_{ab} is known as the lag distance between points a and b, and in general there is a subset of points in a sample set that are a given lag distance apart.

Lag distances often are applied with an associated *lag tolerance*. A lag tolerance

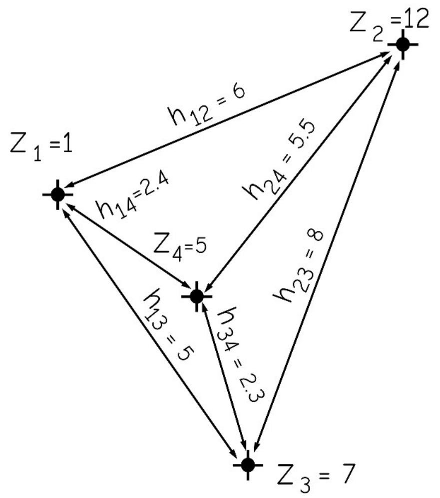


Figure 12-16: Lag distances, used in calculating semivariances for kriging.

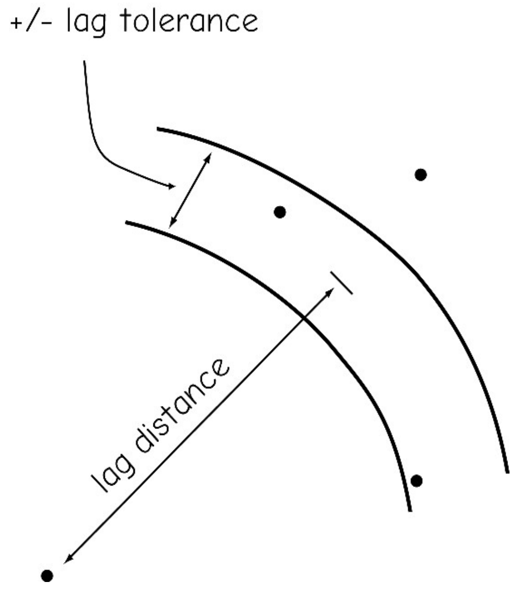


Figure 12-17: A lag tolerance defines a range for grouping samples. The grouping aids estimation of spatial covariance.

defines a small range that is "close enough" to a lag distance (Figure 12-17). A lag tolerance is required because the individual lag distances typically are not repeated in the sample data. Most or all distances between sample points are unique, so there is little or no replication with which to calculate the variability at each lag. Some distances may be quite similar, but distances usually will differ in the smallest decimal places. A lag tolerance circumvents this problem.

The lag tolerance defines when distances are similar enough to be grouped in spatial covariance calculations. For example, we may wish to calculate the semivariance for points that are 112 meters apart. If we are inflexible and only use point pairs that are exactly 112 meters apart (within the precision of our measurement system), we may have only a few, or perhaps even no points that meet this strict criterion. By allowing a tolerance, distances that are plus or minus that tolerance from the given lag distance can be used to calculate a spatial variability. For example, we might set a tolerance for h of 10 units. Any pair of points between 102 and 122 units apart are used to calculate an index of spatial covariance for the lag distance $h = 112$.

Geostatistical prediction uses the key concept of a *semivariance* to represent spatial covariance. A semivariance is the variance based on nearby samples, and it is defined mathematically as:

$$\gamma(h) = 1/2n * \sum (Z_a - Z_b)^2 \qquad (12.7)$$

where Z_a is the variable measured at one point, Z_b is the variable measured at another point h distance away, and n is the number of pairs that are approximately the distance h apart.

The semivariance at a given lag distance is a measure of spatial autocorrelation at that distance. Note that when nearby points (small h) are similar, the difference $(Z_a - Z_b)$ is small, and so the semivariance is small. High spatial autocorrelation means points near each other have similar Z values.

The semivariance may be calculated for any h. For example, when $h=1$, the semivariance $\gamma(h)$ may be equal to 0.3; when $h=2$, then $\gamma(h)$ may be 0.5; when $h=3$, then $\gamma(h)$ may be 0.8. We may calculate a semivariance provided there are sufficient point pairs that are h distance apart to give a good estimate.

We may plot the semivariance over a range of lag distances (Figure 12-18), and this plot is known as a *variogram* or *semivariogram*. A variogram summarizes the spatial autocorrelation of a variable. Note that the semivariance is usually small at small lag distances, and increases to a plateau as the lag distance h increases. This is the typical form of a variogram. The *nugget* is the initial semivariance when the autocorrelation typically is highest. The nugget is shown at the left of the diagram in Figure 12-18, the semivariance at a lag distance of zero. This is the intercept of the variogram.

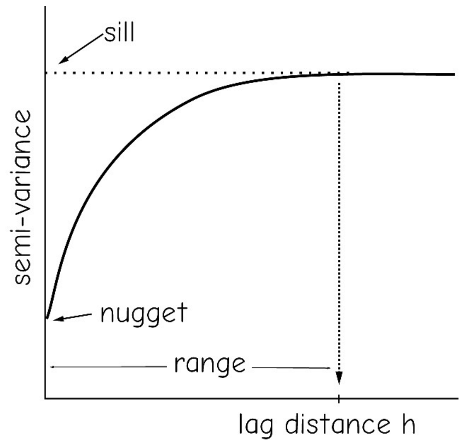


Figure 12-18: An idealized variogram, with the nugget, sill, and range identified.

The *sill* is the point at which the variogram levels off. This is the “background” variance, and may be thought of as the inherent variation when there is little autocorrelation. The *range* is the lag distance at which the sill is reached. The nugget, sill, and range will differ among spatial variables.

A set of sample points is used to estimate the shape of the variogram. First, a set of lag distances h_1, h_2, h_3, etc., are defined; each distance signifies a given lag distance, plus or minus the lag tolerance. The semivariance is then calculated for each lag distance. An example is shown in Figure 12-19. Remember, each of these points is calculated from equation 12.7 for a given lag distance.

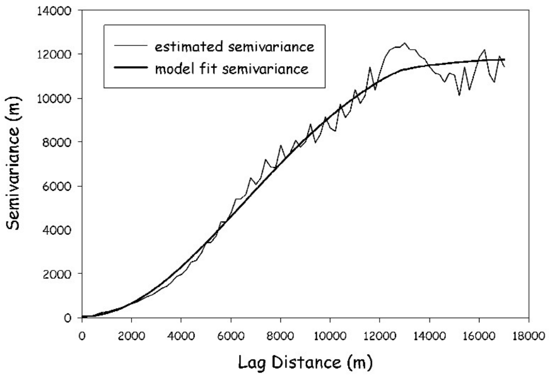


Figure 12-19: A variogram, a plot of calculated and fit semivariance vs. lag distance.

A line may then be fit through the set of semivariance points, and the variogram estimated. This line is sometimes called the variogram model.

Spatial prediction is among the most important applications of the variogram model (Figure 12-20). There are many variations and types of kriging models, but the simplest and most commonly applied rely on the variogram to estimate “optimal” weights for prediction. These weights are used to estimate values at unknown locations by:

$$Q = \sum_{j=1}^{n} w_j \cdot v \qquad (12.8)$$

where Q is the estimated value at an unmeasured point, w_j are weights for each sample j, and v is the known value at sample point j.

Weights are optimal in the sense that they minimize the error in a prediction, and they are unbiased, given a specific data set and model. The calculation of optimal weights requires some rather involved mathematics, beyond our present scope, but is described in great detail in references listed at the end of this chapter.

Estimating each w_j involves a constrained minimization process. A set of equations may be written that expresses the errors as the differences between our measured values and the predicted values by a function of a set of unknown weights. This set of equations is solved under the constraints that the weights sum to zero and the error variance is minimized. The solution involves calculating the expected values of covariances between points according to a variogram model, for example, by fitting a smooth relationship between the observed semivariogram points, as shown in Figure 12-19. The covariances are a function of the specific lag distances observed in the sample, and are used to solve for the optimal set of weights in equation 12.8.

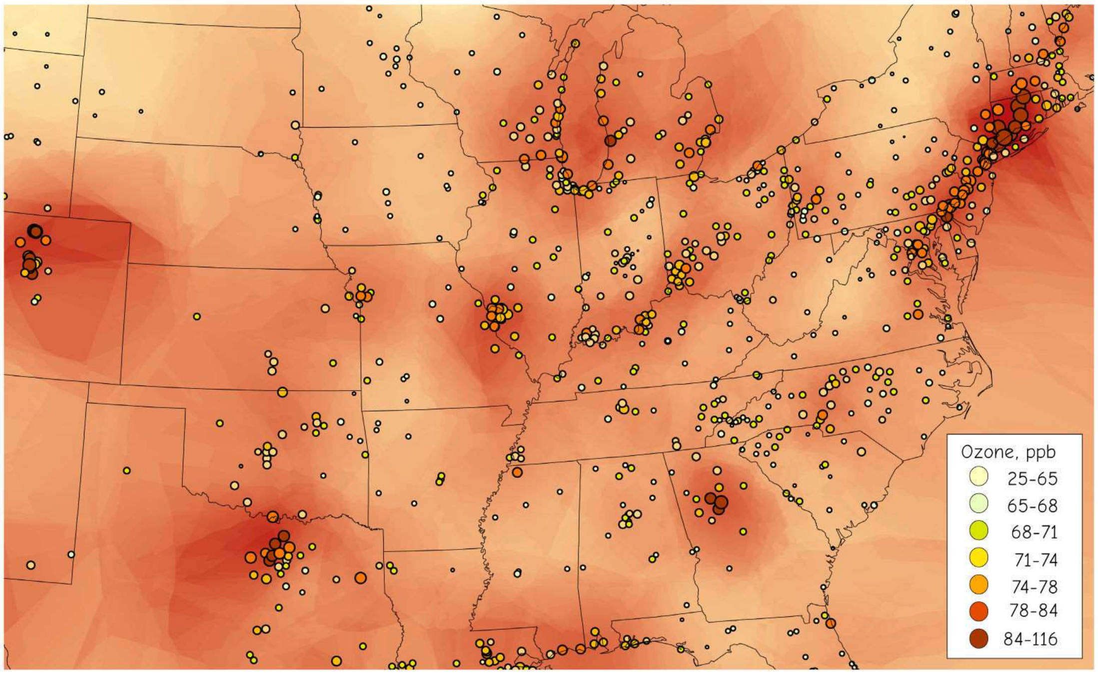


Figure 12-20: Sample points and predicted values from an application of kriging to the ozone data. There are many alternative forms of fitting a kriging model, chosen based on goodness of fit of a variogram model, and on a sample determined interpolation distance.

As stated earlier, kriging is similar to IDW interpolation in that a weighted average is calculated. However, kriging uses the minimum variance method to calculate the weights, rather than applying some arbitrary and perhaps more imprecise weighting scheme as with IDW.

Co-kriging is an extension of kriging that includes the measurement of a separate, correlated variable at the sample locations in addition to the variable of interest. There may be an easily measured secondary variable that is to some extent related to the primary variable, but that is easier or less expensive to measure. In many analyses, temperature might be a primary variable and elevation a secondary variable. Co-kriging exploits the covariance between the primary and secondary variables to improve our estimate of the primary variable. Co-kriging is similar in motivation to kriging in that a set of optimal weights is estimated, but with co-kriging there are weights for both the primary and secondary variables.

Spatial prediction with kriging, co-kriging, and other geostatistical methods can be a complex and nuanced process. There is a wide range of models that may be fit, and these in part depend on the characteristics of the data. Different data characteristics indicate particular modeling methods or model forms, for example, if there are trends in the data, or directional differences in the variance. These considerations are beyond the scope of our present discussion, and the interested reader is referred to more complete treatments, such as Isaaks and Srivastava (1989) or McKillup and Dyar (2010), listed under suggested reading at the end of this chapter.

There are more advanced spatial prediction methods, and spatial estimation is an active area of research, with more complex techniques such as spatial Bayesian estimation and space-time models. These topics are beyond an introductory course, more appropriately treated in more advanced courses and texts.

Prediction Accuracy

We often need to characterize the accuracy of our spatial estimations. This helps us choose the best model and place limits on model application. Model assessment is a well-developed field, and will not be thoroughly reviewed here, but a few main concepts are introduced.

Accuracy is measured at *assessment points*, locations where we know both the true value and the estimated values for a variable.We often describe a sample set with n points, with estimated or interpolated values at any ith point denoted by P_i, and the true or observed value at the point denoted by O_i. Each assessment point provides an error estimate:

$$e_i = P_i - O_i \qquad (12.9)$$

There are several metrics that are commonly used to characterize aggregate error, perhaps chief among them the *root mean squared error*:

$$RMSE = \left[\left(\sum_{i=1}^{n} e_i^2\right)/n\right]^{0.5} \qquad (12.10)$$

Error values are squared to remove the sign effect, and then the square root taken on the sum to return to the measured unit scale, instead of a squared unit scale. Predictions either above or below the observed values are generally considered to be considered equally bad, and the error is averaged over all samples. However, squaring the errors magnifies the influence of outliers, extremely large positive or negative errors, so some argue that this is an overly pessimistic estimate of error, or at least when there are large outliers.

The *mean absolute error* is an alternative error metric, less often used but less sensitive to outliers than the RMSE. The MAE is defined as:

$$MAE = \left[\left(\sum_{i=1}^{n} |e_i|\right)/n\right] \qquad (12.11)$$

It substitutes the absolute value operation for the squaring/square root operations and so is less sensitive to outliers, but otherwise is quite similar to the RMSE.

Another accuracy metric is the *mean bias error*:

$$MBE = \left[\left(\sum_{i=1}^{n} e_i\right)/n\right] \qquad (12.12)$$

MBE measures the average bias in the predictions, the amount by which, on average, an estimated surface over- or underpredicts the true values. MBE conveys useful information overall, but provides little information on the magnitude of individual errors and should be used in conjunction with RMSE, or preferably, MAE.

Overall measures of agreement between an estimated and true surface have been proposed, including Willmott's index of agreement:

$$d = 1 - \frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (|P_i'| + |O_i'|)^2} \quad (12.13)$$

with

$$P_i' = P_i - \bar{O} \quad (12.14)$$

and

$$O_i' = O_i - \bar{O} \quad (12.15)$$

Primary citations of these and other accuracy metrics are provided at the end of this chapter, and in the considerable literature on interpolation and spatial estimation.

Assessing the accuracy of an interpolated surface requires we collect both observed and predicted values at a set of points. In an ideal assessment, these would be independent of the samples we use to estimate the surface, but this is rarely possible. Samples are often expensive, difficult to collect, and sparse, and most interpolated surfaces would benefit from additional sampling. If each new sample can materially improve our interpolation, we are hard-pressed to hold them in reserve for an accuracy assessment. We are tempted to use most or all of our samples while interpolating, and leave few or none for an accuracy assessment.

Exact interpolators are particularly vexing. As you might recall, Thiessen polygons, inverse distance weighted, and some spline interpolators have zero error at all sample points by definition, because they are formulated to exactly return the observed values at the fitted points. One might think that we must hold a set of points in reserve in order to get a true estimate of the interpolator accuracy.

There are related techniques, known variously as leave-one-out, bootstrapping, or cross-validation, which addresses both the undersampling and robust accuracy estimation requirements. Bootstrapping involves fitting the surface as many times as there are sample points, each time withholding one of the points. We fit the surface the first time, withholding the first point. We can then subtract the withheld measured value (O_1) to the interpolated value (P_1), and obtain one estimate of the error. We then repeat this process for the rest of the sample points. For n samples, we fit the surface n times. We can then compare the withheld point's true value, O_i, to the fit value P_i, giving us n error values, e_i. We can then apply equations 12.10 through 12.15 to characterize the accuracy of our fits.

Unfortunately, most surface interpolation tools in GIS do not support bootstrapping or similar validation methods. This is unfortunate, doubly so because they typically provide only the RMSE value, and then only without bootstrapping or other cross-validation, and for inexact interpolators. RMSE estimated from the fit points without bootstrapping may well give an optimistic estimate of accuracy, particularly when sample size is small, and should not be accepted in lieu of a bootstrap or similar validation. This should be remedied, if not within the specific GIS software used in fitting, then by exporting the sample data to a statistically oriented surface fitting system; for example, the open source statistical package R.

Core Area Mapping

Core area mapping is another common and useful spatial analysis tool. A *core area* is a primary area of influence or activity for an organism, object, or resource of interest. Detectives may wish to map a series of burglaries to uncover clustering or patterns in occurrence. Wildlife managers may wish to map the home range of an endangered organism, or a business owner the home locations of her customers.

Core area mapping typically involves identifying area features (polygons, raster areas, or volumes) from a set of point or line observations. Individual burglaries, for example, are recorded as point locations, perhaps tagged to the address or building where they occurred. These points may be used to define a polygon by one of several core area mapping techniques. In this way, the core area is a higher dimensional spatial object (area) that is defined from a set of lower dimensional objects (points or lines). This core area represents some central or important region where features occur frequently, in this example, burglaries. Additional resources may be focused on this core area, such as increased patrols or surveillance.

Core area mapping is commonly used. Perhaps the most frequent applications to date have involved analysis of patterns of human activity, particularly crime occurrence, as illustrated in the previous example. In addition, plant and animal species densities are often analyzed and summarized using these methods, particularly when the organism is highly valued or endangered. Resource managers record organism occurrences in the field, perhaps using GPS or other spatial positioning technologies. These observations may be combined and abundance patterns are analyzed after a sufficient number of observations has been gathered. Core areas may be identified, and key habitat conditions or requirements inferred. These may guide management actions such as the protection of areas with a high concentration of endangered species and the enhancement of other areas by adding key habitat requirements.

Mean Center and Mean Circle

The *mean center* and associated *mean circle* are perhaps the simplest and most obvious measure of a central location and a core area. The mean center is simply the average X and Y coordinates of the sample points. Each sample point has an associated pair of coordinates. These may be summed and the average calculated, and this mean point identified as the center of the core area.

Mean circles may be associated with the mean center to define a core area (Figure 12-21). The mean circles are defined by a radius measured from the mean center. The mean circle radius is commonly the distance to the farthest sample point, the average distance from the mean center to the set of sample points, or some other statistical measures based on the variance of the distance to sample points. These distances may be calculated easily from the sample X and Y coordinates, first by calculating the mean, and then by applying the general formulas to

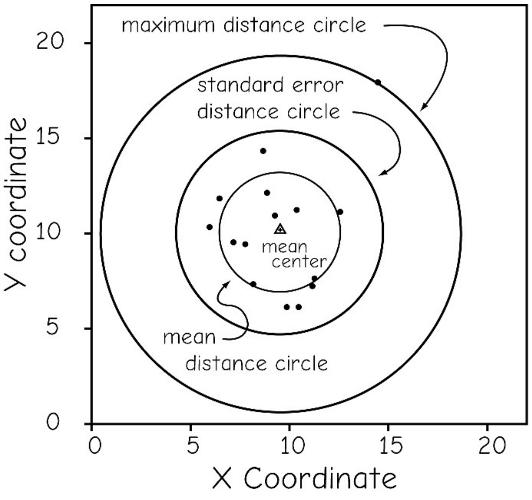


Figure 12-21: An example of a mean center and corresponding mean circles for a set of sample points.

calculate distance from sample points to the mean center. The largest distance, average distance, or the standard deviation of the distance from points to the center then may be determined.

Mean circles have the advantages of simplicity and ease of construction, but they assume a uniformly circular shape for the core area. Some measures of mean center may be biased by extreme points; for example, the maximum distance circle in Figure 12-21. Note that the outlier near X = 15 and Y = 17.5 results in a large maximum distance circle. This circle contains substantial area with no points nearby, and it is probably an overestimation of the core area. It is not clear that the mean distance or standard error distance circles are better at defining a core area. The core areas defined by these measures may be appropriate for some applications, but they are often too small in others. Some multiple of the mean distance or standard error may be chosen based on statistical assumptions, or past experience. For example, if we assume the samples follow a random normal distribution, then a core area defined by a circle approximately 1.8 times the standard error distance should contain 68% of the data. Previous experience may help; for example, one might know that in a particular region, 90% or more of a wolf pack core area is within 10.8 km of a mean center.

In many cases, circular core areas are suboptimal because many variables are known to exhibit nonregular shapes, and a circular core area is identified when using the mean center / mean circle methods. While mean circle methods are often used in exploratory data analyses, other methods have been developed to more effectively identify irregularly shaped core areas.

Convex Hulls

Convex hulls, also known as minimum convex polygons, are perhaps the simplest way to identify core areas with irregular shapes. A *convex hull* is the smallest polygon created by edges (lines) that completely enclose a set of points, and for which all exterior angles between edges are greater than or equal to 180 degrees (Figure 12-22). An exterior angle is measured from

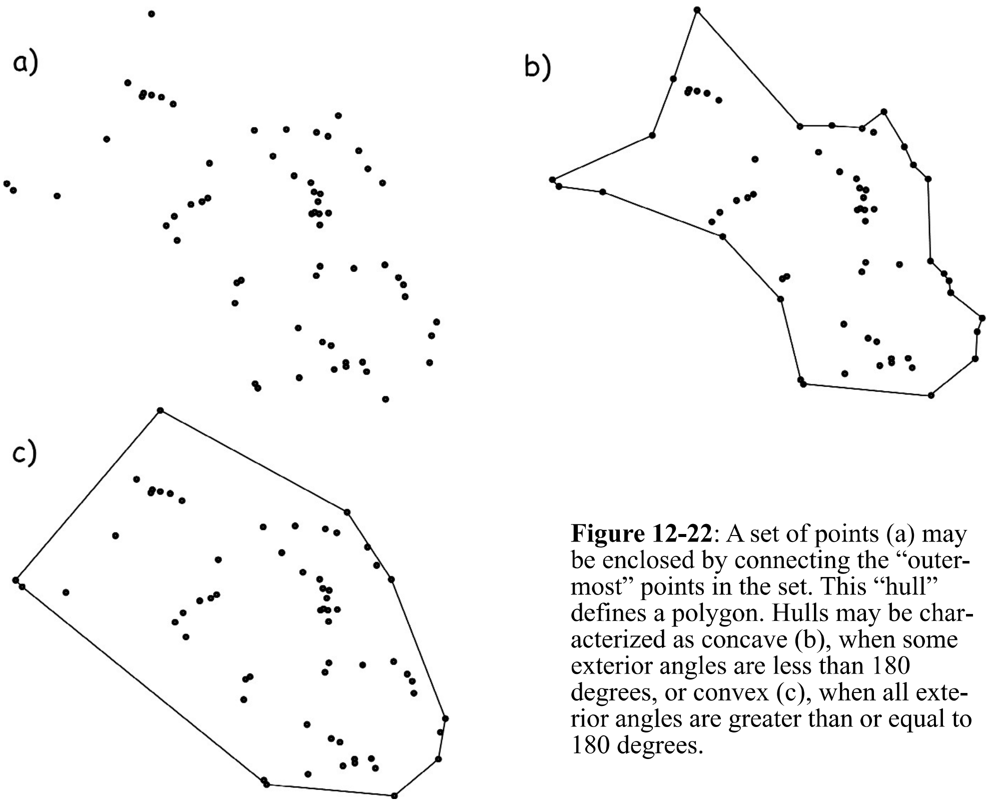


Figure 12-22: A set of points (a) may be enclosed by connecting the "outermost" points in the set. This "hull" defines a polygon. Hulls may be characterized as concave (b), when some exterior angles are less than 180 degrees, or convex (c), when all exterior angles are greater than or equal to 180 degrees.

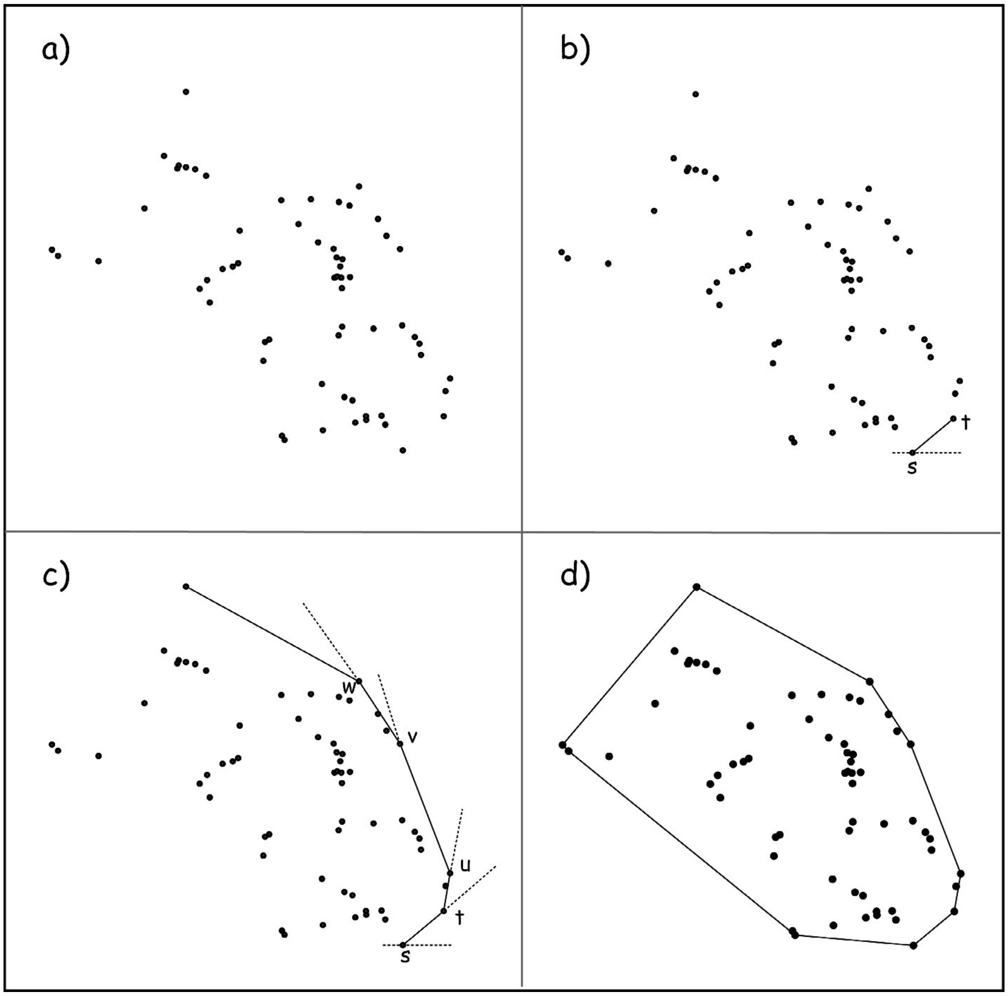


Figure 12-23: The convex hull for a set of points (a) may be calculated from a sweep algorithm. Starting with an extreme point such as s (b), successive minimum deflection angles are selected (c) until the starting point is reached (d).

one edge or side to another through the region "outside" of a polygon. Squares, triangles, and regular pentagons are all examples of convex hulls, while stars and crosses are examples of nonconvex hulls. While these geometric figures have regular shapes, most convex hulls derived from sampled points will not.

Convex hulls are often considered a natural bounding area for a set of points. This assertion is accepted by most analysts when there are no outlying data points, far removed from the rest. When outliers are present, the convex hull will often be unreasonably large.

Convex hulls are widely used because they are simple to develop and interpret, and there is little or no subjectivity in their application. The shape of the convex polygon is determined solely by the arrangement of the sample points, and not by controlling parameters that must be specified by the human applying the method. They represent the irregular shapes common to most sampling.

A convex hull may be easily created with a "sweep" algorithm applied to a set of sample points (Figure 12-23a). These are the locations of the events of interest, for example, observations of a rare animal or crime locations. An extreme point is identified from the set, usually the sample with the largest or smallest X or Y coordi-

nate (point s in Figure 12-23b). The angles of deflection from the current point to all other points are calculated, and the smallest positive clockwise or counter clockwise angle and corresponding point are identified (point t in Figure 12-23b). This point is the next in the convex hull, and becomes the starting point for the next calculation. This process is repeated until the starting point is reached (Figure 12-23c and d).

Convex hulls are often considered a natural bounding area for a set of points. However, convex hulls often ignore clustering in the data. A dense cluster of points in an interior region does not influence the shape of the core area. We lose much of the information on density or frequency of occurrence in the interior region of the bounding polygon. Algorithms defining optimum concave polygons have been developed, generally fitting convex hulls to successive subsets of bounding points, and discarding outlier points, or areas defined by the outlying points. One such method is described next.

Characteristic Hull Polygons

An alternative to convex hulls has been developed, known as characteristic hull polygons (CHP). A Delaunay triangulation is created among the sampled points, the same method described in Chapter 2 when developing a triangulated irregular network. A set of minimum spanning triangles is created, and this set of triangles winnowed to remove a largest area or longest perimeter subset. Figure 12-24a shows a set of sample points and the resulting convex hull, while Figure 12-24b shows the Delaunay triangulation for the same set of points. In this example, the top 5% of polygons with the longest perimeter have been discarded, and the remaining shaded to represent a core area. This reduces the influence of distant points and allows for "holes" embedded within a core area, two advantages over convex hulls. One must choose whether to use area, perimeter, or some other metric of size, so the resultant CHP size and shape depend on the threshold value chosen, for example, 5 vs. 10

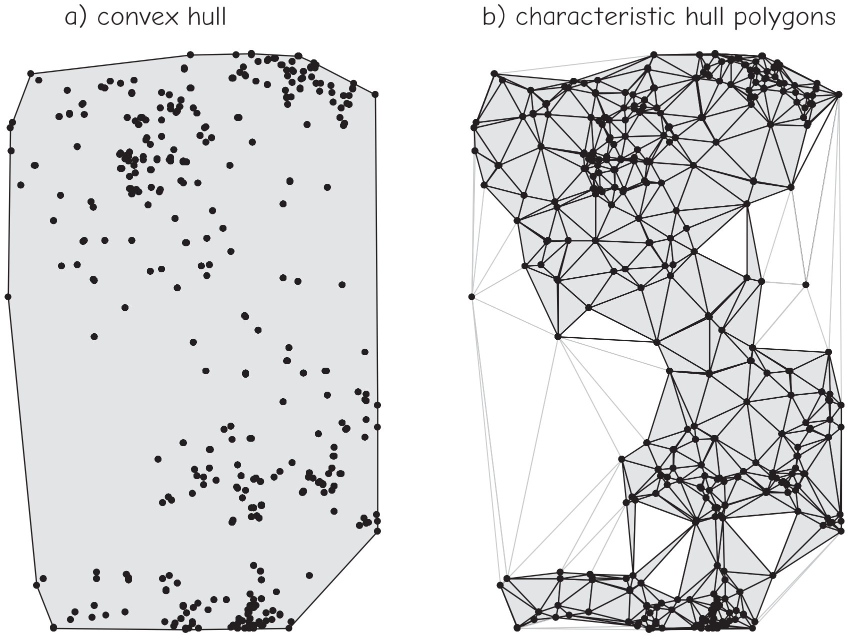


Figure 12-24: An example of a core area defined by a convex hull (a), and characteristic hull polygons (CHP, in b). The shaded area is offered as a core area, with a smaller, higher sample density, and arguably more accurate area identified by the CHP.

largest polygons; however, the method is easy to apply and arguably provides a better estimate of core areas when compared to a convex hull, particularly when outliers are frequent.

Kernel Mapping

Kernel mapping uses a set of sample locations to estimate a continuous density surface. Kernel mapping is widely applied because it is mathematically flexible, relatively easy to implement, may be robust to outliers, readily incorporates clustered samples, can represent irregular shaped core areas, and is often statistically based.

Kernel mapping is based on a density distribution that is assumed for each sample point. These density distributions are placed over the sample plane, one for each observation point, and vertically added to determine the composite density from the sample. This composite density may be used to identify a core area, selecting the densest areas first.

An example will help illustrate these ideas and the process of kernel mapping. Consider samples to detect the density of defects in a tile floor. Each tile is 0.5 in across. We count the number of defects per tile, beginning at one edge of the tile mosaic. We will show the samples collected along a line, but the process and principles are similar in two dimensions.

Figure 12-25 shows the results of a sampling along a line segment. One defect, or fault, is found on a tile located 2 in from the start, and it is represented by a rectangle two units tall. Each fault represents a density of two units, because each tile is 0.5 in across — hence 1/0.5 = 2 faults/in. We observe two faults at 2.5 in (four faults/in), one at 3.5 in, and additional observations until our last fault observed at 12.5 in. Note that the density is in the form of rectangles that are "stacked" two units high for each fault observed for a tile.

Note two things about the density estimates. First, we assume a characteristic shape for the density derived from each observation. In Figure 12-25, we assume the shape of a rectangle for each observation, with a uniform density across the tile.

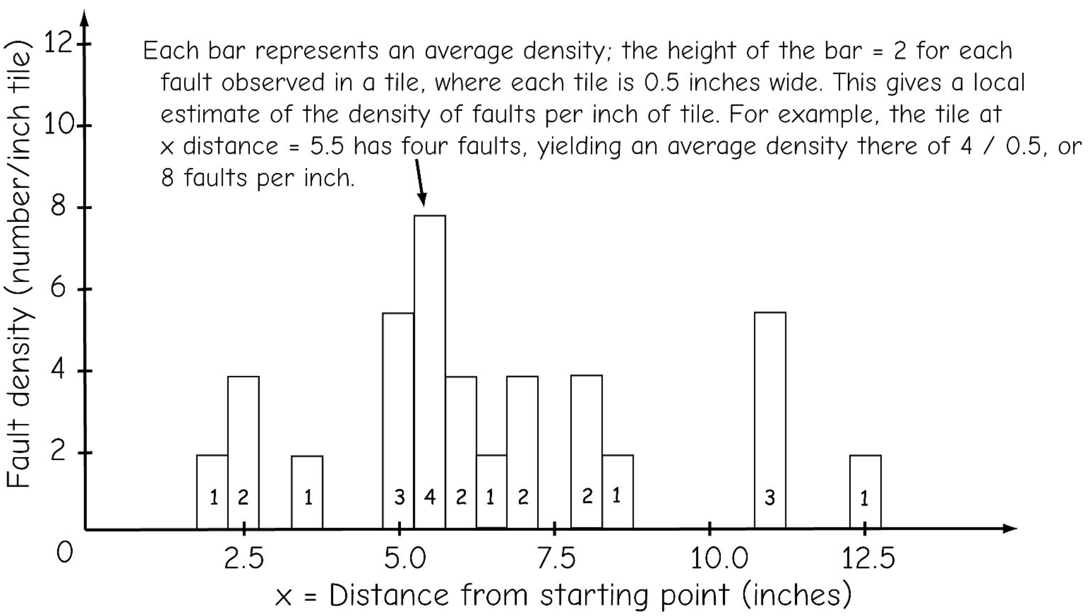


Figure 12-25: Kernel mapping is based on the concept of a distribution of observations in space. Each observation contributes information about our cumulative distribution and the observations are combined to approximate our cumulative distribution.

This may not be true, but in our case we are using a discrete sample, and so it is a valid approximation. In general, this shape is called a density distribution. This characteristic shape (density distribution) is then placed for each observed sample; for example, note that there is a rectangle placed for each defect we observe at a distance from the starting point in Figure 12-25.

Second, note that the shapes (density distributions) are added vertically in areas where they coincide, as shown in Figure 12-25. In our example, rectangles are stacked. With more complex, mathematically-defined density distributions, the values are added over each point. The cumulative density distribution is the sum of the distributions associated with each sample.

Density distributions typically are not squares or other geometric figures, but rather symmetric shapes such as parabolas, Gaussian curves, or otherwise smoothly varying surfaces about a center point. These shapes can be mathematically defined and specified for each sample point. For example, a general Gaussian curve for one variable has the form:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-x_o)^2}{2\sigma^2}} \quad (12.16)$$

where x_o is the sample location and σ is a scaling constant. This is a symmetric function about x_o, meaning the function is a mirror image reflected across both sides of the point x_o (Figure 12-26). Note that the density distribution in the figure reaches a peak at x_o, and the area under the curve is typically equal to one. The formula is often written with $\sigma^2 = 1$, or may be scaled by dividing by a value h, so that it appears as:

$$f(x)_h = \frac{1}{h\sqrt{2\pi}} \cdot e^{\frac{-(x-x_o)^2}{2h^2}} \quad (12.17)$$

The value h is also known as a *bandwidth parameter*, and is described in the next few paragraphs.

Many functional forms can be used to represent the kernel densities. Typically, these shapes are "bumps" in that they smoothly rise to a peak and then descend to near zero. Different forms of the kernel density function may have characteristic shapes — how fast they reach the peak, how pointed the peak becomes, and how quickly they return to values near zero at points more distant from the peak.

The composite density distribution is created by "stacking" our individual density distributions from the set of observations (Figure 12-27a and b). Density distributions may be plotted for each observation; for example, two of many observations are shown in Figure 12-27a. Each point yields a smooth "bump" centered on the observation.

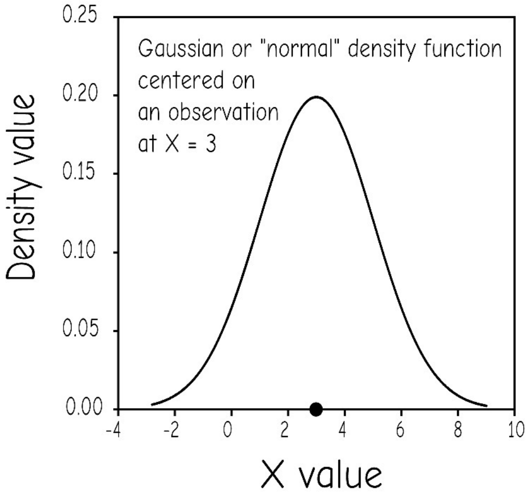


Figure 12-26: A density distribution is assumed, and plotted for each observation. Here an observation at X = 3 (plotted dot) generates a bell-shaped curve centered on the observation.

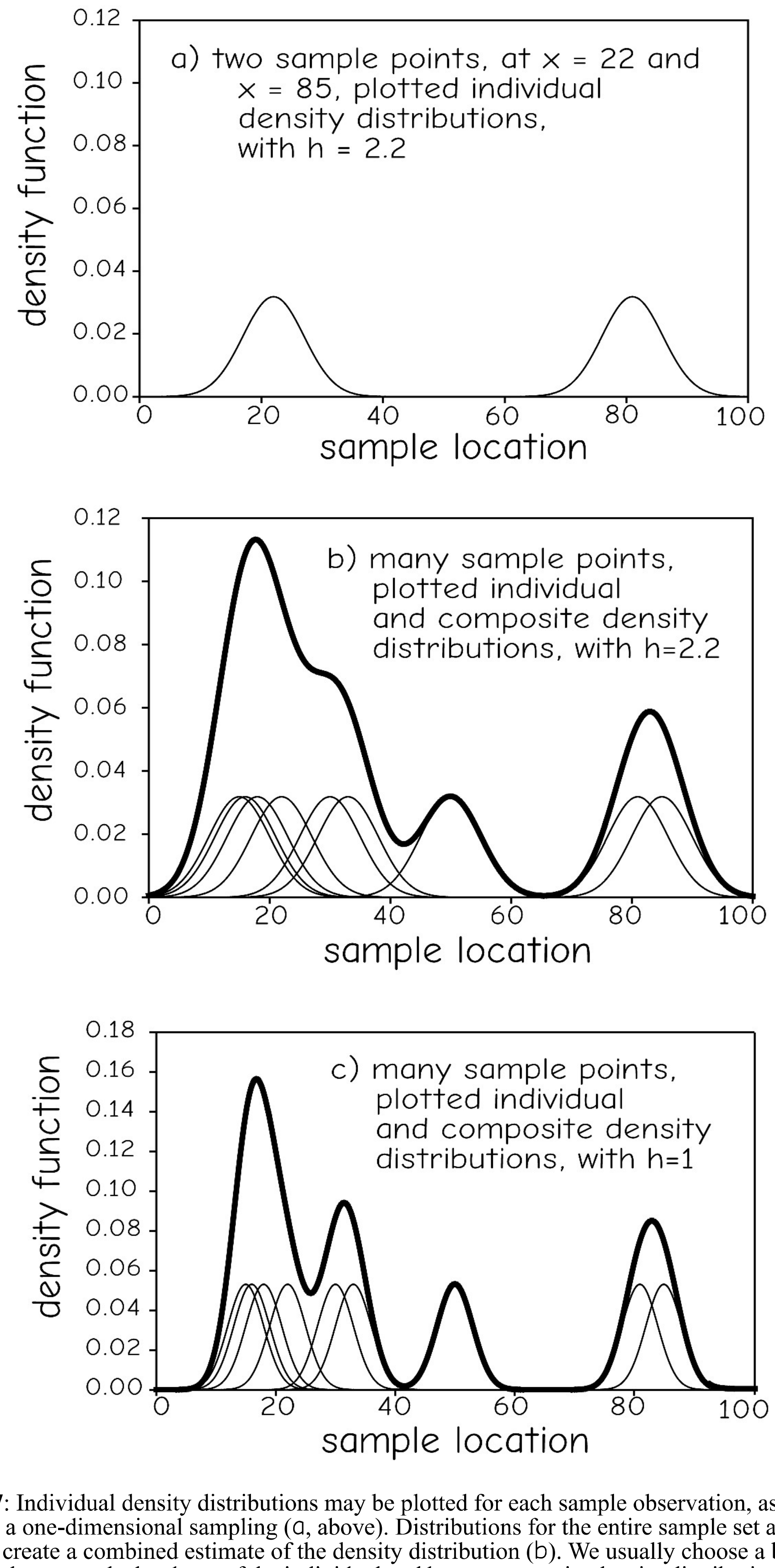


Figure 12-27: Individual density distributions may be plotted for each sample observation, as shown for two points in a one-dimensional sampling (a, above). Distributions for the entire sample set are plotted, and added to create a combined estimate of the density distribution (b). We usually choose a bandwidth parameter, h, that controls the shape of the individual and hence composite density distributions. Narrower bandwidths result in higher and narrower peaks (c).

When all observed points are plotted, there is a commensurately large number of small, overlapping bumps, as shown by the thin lines in Figure 12-27b. These may then be summed vertically to create the cumulative density distribution, shown by the thick line in Figure 12-27b.

We often choose bandwidth parameters, symbolized by h, that define the "spread" or width of the individual density distributions (Figure 12-27c and Figure 12-28). Perhaps the simplest way to understand the bandwidth is to think of the binning interval in our example in Figure 12-25. There, we counted tile defects for each 0.5 in tile, and plotted a rectangle corresponding to the resultant fault density. Our bandwidth was set at 0.5 in. We just as well could use a bandwidth of 1 in, counting the number of defects per two tiles (1 in), along our sampling line. This would give a related, but slightly different estimate of the density distribution of defects along our sampling line. As shown in the right panel of Figure 12-28, the estimated density distribution for the first 7 in of our sampled line is less "peaked" or "spikey." Although the same sample set is used to estimate both density distributions, each observation is spread across a broader interval when we choose a larger bandwidth.

We observe the same change in the peakedness when we change the bandwidth for continuous density distributions, such as the Gaussian distribution shown in Figure 12-27 and equation 12.16. A sample is plotted using a Gaussian density function for each observation and a bandwidth of h = 2.2 in Figure 12-27b. Reducing the bandwidth to h = 1 narrows the shape for each individual sample and results in higher, narrower, more peaked shapes in the cumulative distribution shown in Figure 12-27c.

Kernel mapping is generally a three-step process, as may be surmised from the preceding discussion. First, we collect samples and the concomitant coordinate locations. Second, we choose a kernel density function. Finally, we choose a bandwidth, h, apply the kernel density distribution, and sum across each sample area to achieve our composite estimate of density.

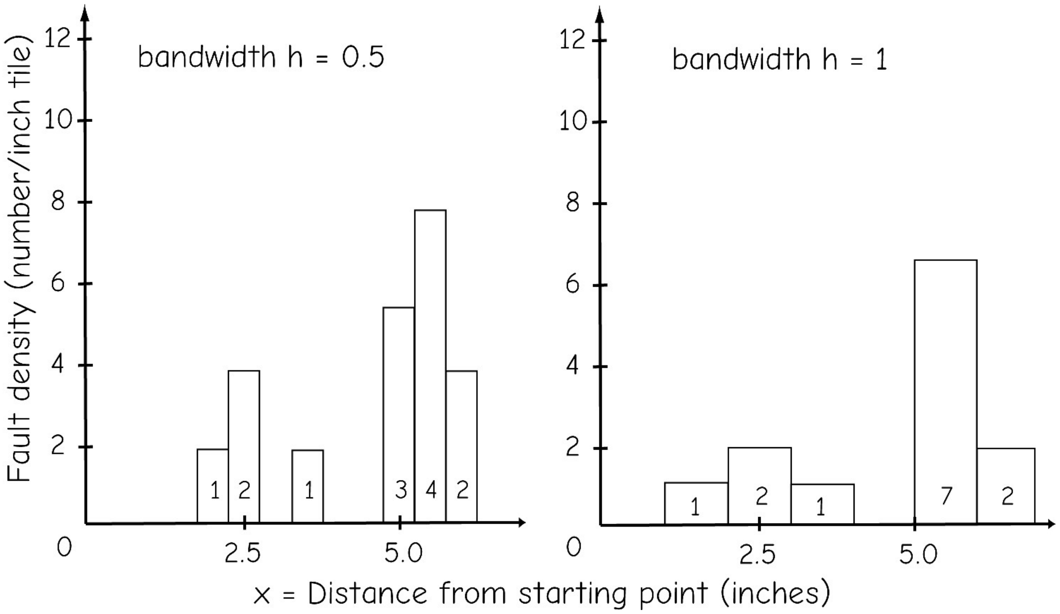


Figure 12-28: We choose both the general form of the density distribution, and a bandwidth parameter that affects the shape of the distribution. Here, the rectangular shape of the fault density becomes broader and shorter as the bandwidth changes from 0. 5 to 1.

Mathematically, this process is summarized by the equation:

$$\lambda(x, y) = \frac{1}{nh^2} \cdot \sum_{i=1}^{n} \frac{K(x_i, y_i)}{h} \quad (12.18)$$

where $\lambda(x,y)$ is the composite density distribution, n is the number of samples, h is the bandwidth, and $K(x_i,y_i)$ is the individual density distribution applied at each sample point i.

An example of kernel density estimation is shown in Figure 12-29. An individual sample point is shown in Figure 12-29a, with a single peak corresponding to the Gaussian density distribution chosen. A more complex shape with multiple peaks occurs when all sample points are plotted, as shown in Figure 12-29b. Individual distributions are summed vertically, resulting in an undulating, complex surface. This surface represents the density or probability of occurrence of the underlying variable, for example, the density of defects in a tile floor, the crime density mapped across a city, or the utilization density for a wolf pack in their home range.

While the choice of bandwidth affects our results, there is no uniformly best method to select the appropriate value for h. One commonly applied method is to plot several density surfaces, one for each of a given h value, and select the h that most closely approximates your perception of the best density. Insights in the distribution and behavior of the data set are often gained by analyzing densities across a range of bandwidth values.

Formulas exist for optimum bandwidths under various conditions. One method for calculating optimum bandwidth has been proposed by Fotheringham et al. (2000) for a Gaussian kernel:

$$h_{opt} = \left[\frac{2}{3n}\right]^{\frac{1}{4}} \sigma \quad (12.19)$$

where h_{opt} is the optimum bandwidth, n is the number of samples, and σ is the standard deviation parameter, unknown, but estimated from the sample.

Numerous formulas exist defining optimum bandwidths, and one is faced with a rather different choice of selecting the cor-

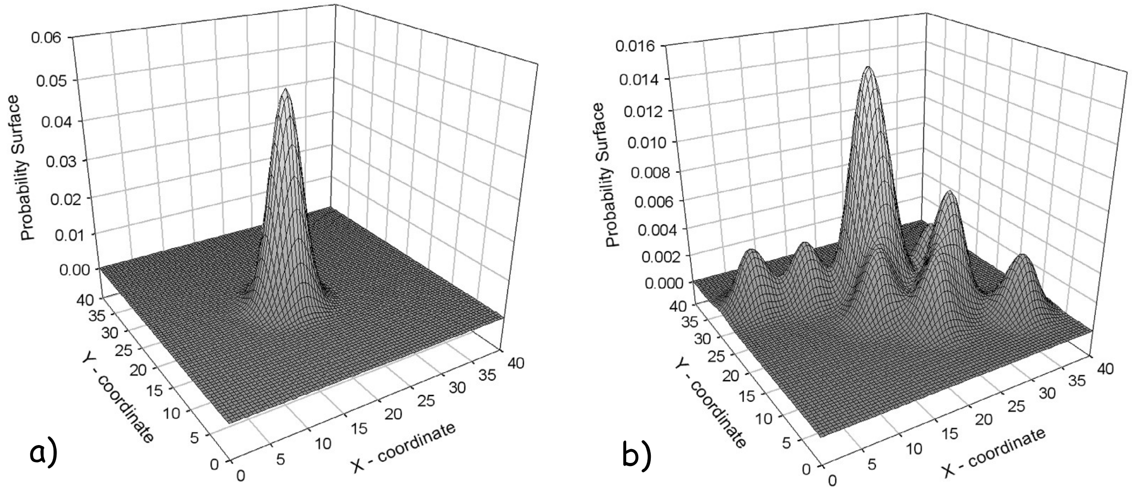


Figure 12-29: Kernels may be used to map density distributions across a two-dimensional surface. Density distributions from individual samples (a) are summed to create a composite estimate of the density surface (b).

rect optimum. The motivations behind various optimum bandwidths are described in the books by Silverman (1986) and by Fotheringham et al. (2000), listed at the end of this chapter.

Core area delineation is a primary use for estimated density distributions. As expected, the identified core areas are dependant on the selected bandwidth. Figure 12-30 shows vertical views of two-dimensional density distributions for optimum (a), below-optimum (b), and above-optimum (c) bandwidth. Darker shades of gray show higher densities, and note the narrower, more concentrated distributions at the lowest bandwidth (b) relative to the largest bandwidth (c). These different bandwidths result in different core area polygons (d through f). Empirical tests and experience guide the choice of best bandwidth.

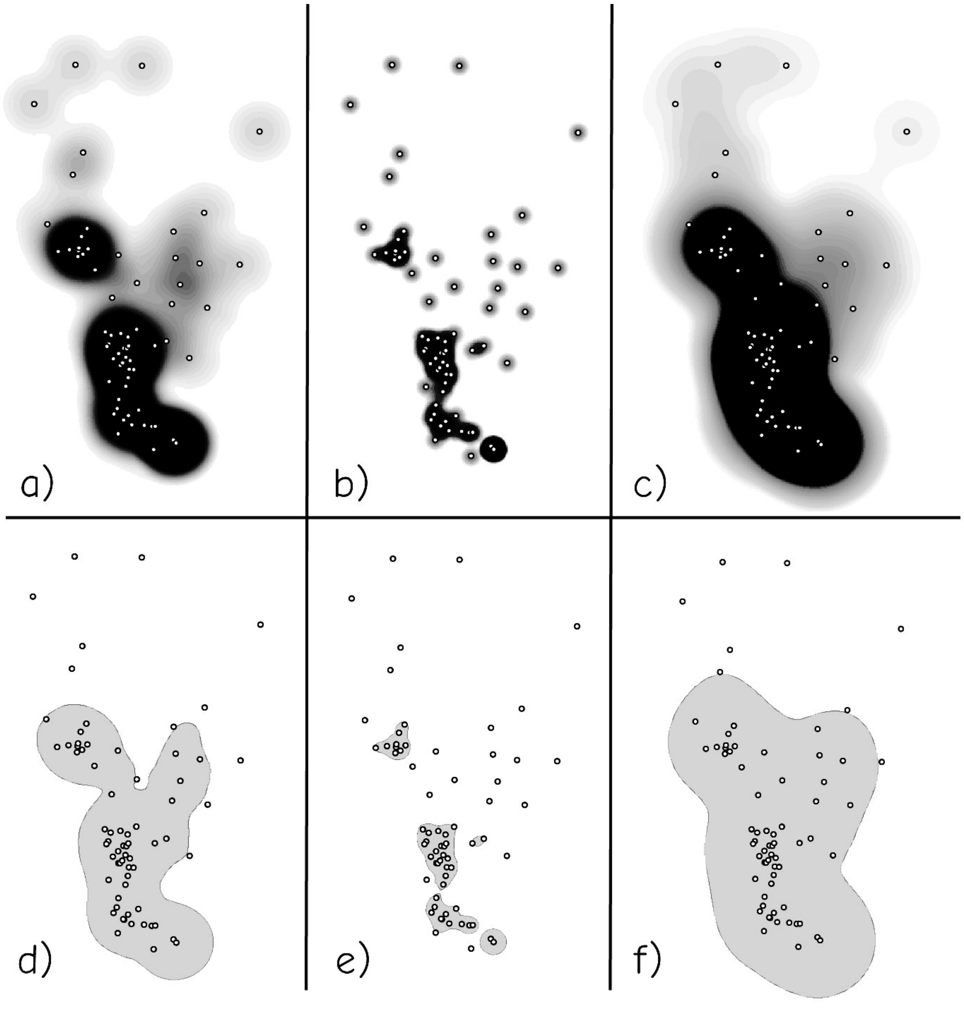


Figure 12-30: Kernel mapping may be used to identify core areas, although the areas that will be defined depend on the method used. Panels a through c show Gaussian density distributions for a sample set under varying bandwidths, while d through f show corresponding core areas encompassing 90% of the density distributions.

Time-Geographic Density Estimation

Density estimators have been developed for space utilization by moving objects, typically animals for home range analysis, although sometimes other objects. An object may be observed periodically through space, for example, when a GNSS is attached to a migrating penguin, and the position relayed to a base station. These positions are often called control points, because they establish the location of the tracked object at a fixed point in time. This sequence of control points defines a path (Figure 12-31).

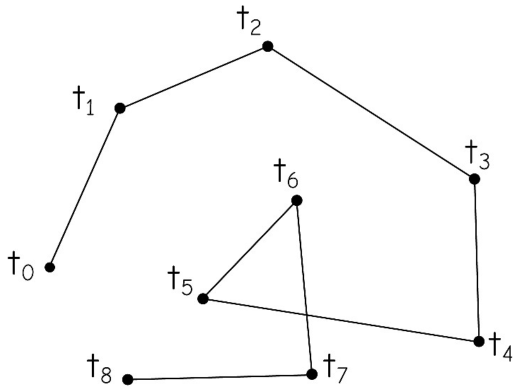


Figure 12-31: A sequence of observations on a moving object, from an initial time (t_0) to a final time (t_8).

While locations between observations cannot be precisely determined, the control points constrain where the penguin might have been, because there is an upper limit on how fast the bird can travel. We may establish a maximum velocity, v, either from previous observations, from the current tracking effort, or from theoretical limits. Time-geographic density estimation (TGDE) combines a sequential set of control points with knowledge about maximum velocity to estimate spatial occurrence probabilities.

TGDE depends on the concept of a geoellipse between two points. If P_i is the control point at time i and P_j the control point at time j, then the geoellipse g_{ij} may be defined as:

$$g_{ij} = \{P|[D(P, P_i) + D((P, P_j) \leq ML)]\} \quad (12.20)$$

where $D(P, P_i)$ is a distance between any point P and the control point P_i, and ML is the maximum distance the object could possibly travel between the successive control points P_i and P_j. ML may be estimated by:

$$ML = (t_j - t_i) \cdot v \quad (12.21)$$

where t_j is the time of observation of control point j, and v is the maximum velocity for the object.

Figure 12-32 illustrates a geoellipse for two control points, P_i and P_j. Note that the distance function need not be Euclidian distance, but it usually is. The tracked object is restricted to have been within the drawn ellipse, provided our estimate of v is valid. The size and shape of the ellipse depend on the distance between the successive control points, the time interval between the observations, and the maximum velocity possible. Successive points near each other relative to the maximum distance, given the time differ-

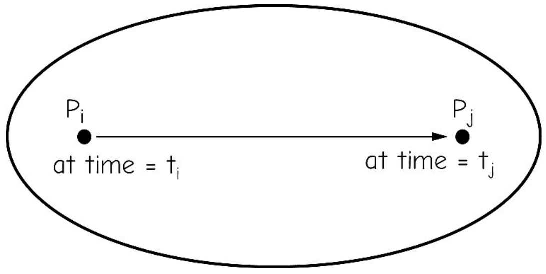


Figure 12-32: An object is estimated to have been contained within an ellipse, given two subsequent control points. The ellipse describes the furthest an object could have reached in any path traveling from P_i and P_j over the time interval t_i to t_j.

ence and maximum velocity, will be enclosed in a nearly circular ellipse, while successive points very near the maximum possible distance will be joined by a long, very narrow ellipse.

Much as when using kernel density functions for estimating a core area, a time-geographic estimate of space use is a composite of many observations. Here, each pair of observations may be considered a density volume, proportional to the probability that the object occupied a location during the time interval (Figure 12-33). A uniform density function is the simplest to understand, implying the object was moving at maximum velocity between the two controlling observations, but along an unknown path within the ellipse. A uniform density function should have a volume equal to the likelihood of occupancy, as with a standard kernel density estimator. For simple shapes such as a three-dimensional uniform probability distribution, the volume is equal to the area times height.

Figure 12-33 shows two points, P_1 observed at time t_1, and P_2 observed at time t_2. Our task is to calculate the area of the elliptic volume that represents the occupancy probability, given our observations. Two paths are shown between the points, one traveling distance d_m and another d_n. These two paths have the same length, by the definition of the bounding ellipsoid, and they are also each equal to the long axis length of the ellipsoid, $2a$. Geometric relationships between the interpoint distances and the dimensions of an ellipse allow us to calculate a and b, two characteristic dimensions, which in turn allow us to calculate the area, πab. This may then be scaled by height to assign an occupation probability (Figure 12-33, lower half).

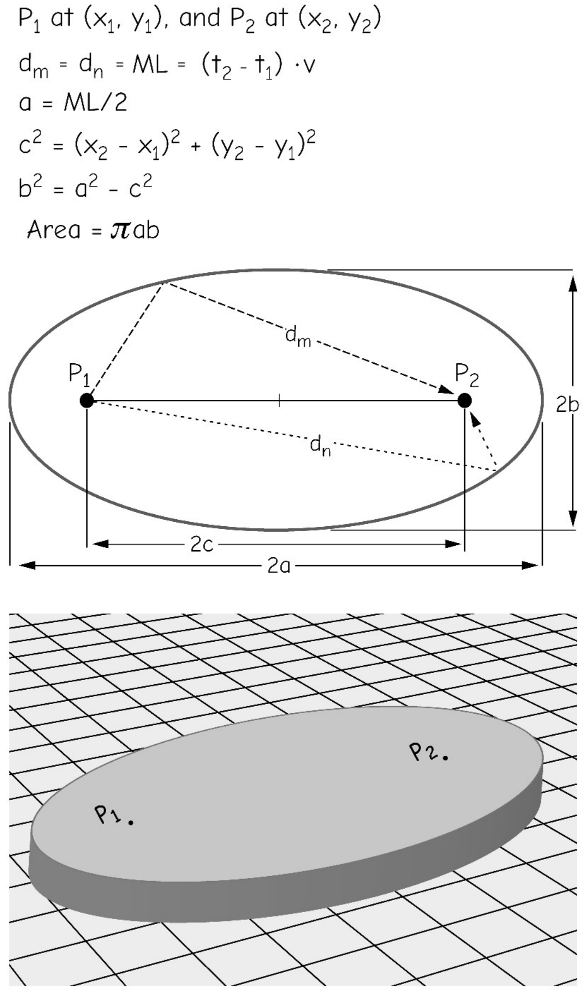


Figure 12-33: The process of calculating a geoellipse density between two sequential points, representing the likelihood of occupation. Here, a uniform probability is assumed across the observations.

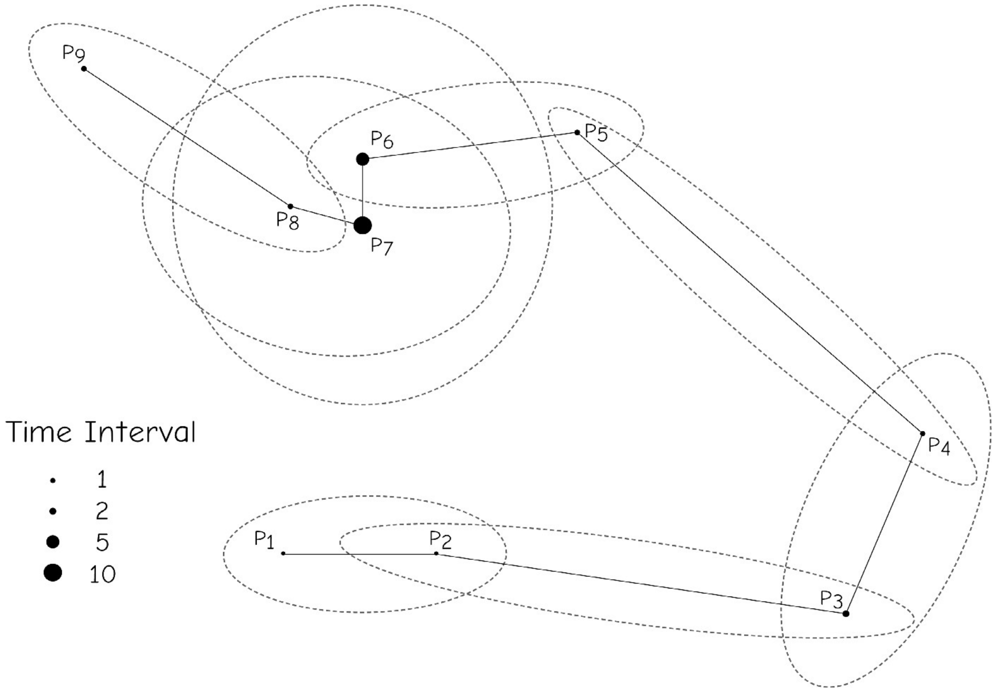


Figure 12-34: An example of the overlapping set of geoellipses used to create a composite density function from a time-geographic data set. The sequence begins with P_1 in the lower left, through P_9 in the upper left. Note that the control point symbol size denotes the time interval since the last control point observation (adapted from Downs et al., 2011).

The process is repeated for overlapping point pairs across the set of observed control points. The next two points in the sequence, P_2 and P_3, are paired, and the density ellipse calculated, summing the densities where geoellipses overlap. The process is repeated for points P_i, P_{i+1} until the last point is reached. Figure 12-34 illustrates the overlapping set of geoellipses from a sample set.

Ellipses may vary in shape, depending on time interval, distance between features, and the maximum velocity (Figure 12-34). Longer time intervals between observations result in larger ellipses, irrespective of the distance between subsequent points. As the interpoint distance approaches the maximum set by the maximum velocity, v, the ellipses become longer and narrower, and reduce to a line when the points are spaced at the maximum possible distance. Conversely, the ellipses approach circles when the time interval between points is long but the observed distance between points is small. This occurs when the object has not moved much, relative to how far it might have moved in the time interval between observations.

The composite time-geographic density function is shown in equation 12.22, where f(x) is the density at any point across a surface; n is the number of observed control points; t_s and t_e are start and end times, respectively; t_i and t_j are consecutive point pairs; v is the maximum velocity and $D(P, P_i)$ is the distance function, as described in equation 12.20. This equation is used for a set of points to estimate the density across space. The numerator sums the weighted

$$f(x) = \frac{1}{(n-1)[(t_e - t_s)\cdot v]^2} \sum_{i=1}^{n-1} H\left[\frac{D(P,P_i) + D(P,P_j)}{(t_i - t_j)\cdot v}\right] \tag{12.22}$$

distance ellipse functions for each pair of sampling points, and the denominator scales this by the maximum distance that may have been travelled during that time interval.

The composite of individual ellipses may result in complex aggregate density volumes. Densities will be highest where points are clustered or near where paths intersect frequently. Sharp edges and sampling artifacts may occur when using a uniform density function, at least until sample size becomes large.

Although Figure 12-33 illustrates a TGDE using a uniform distribution function, other functions may be used. One form assumes the likelihood of occupation decreases linearly with the distance from the line connecting two subsequent control points. This is often called a linear decay function, because the occupation likelihood is assumed to decrease linearly with distance. A more rapid or less rapid decrease with distance may be represented by other functions.

The composite time-geographic density estimate in Figure 12-35 illustrates a space–time path and a linear decay function applied to associate a probable occupancy area for the path. Panel a shows the control points for a path, with the time between successive

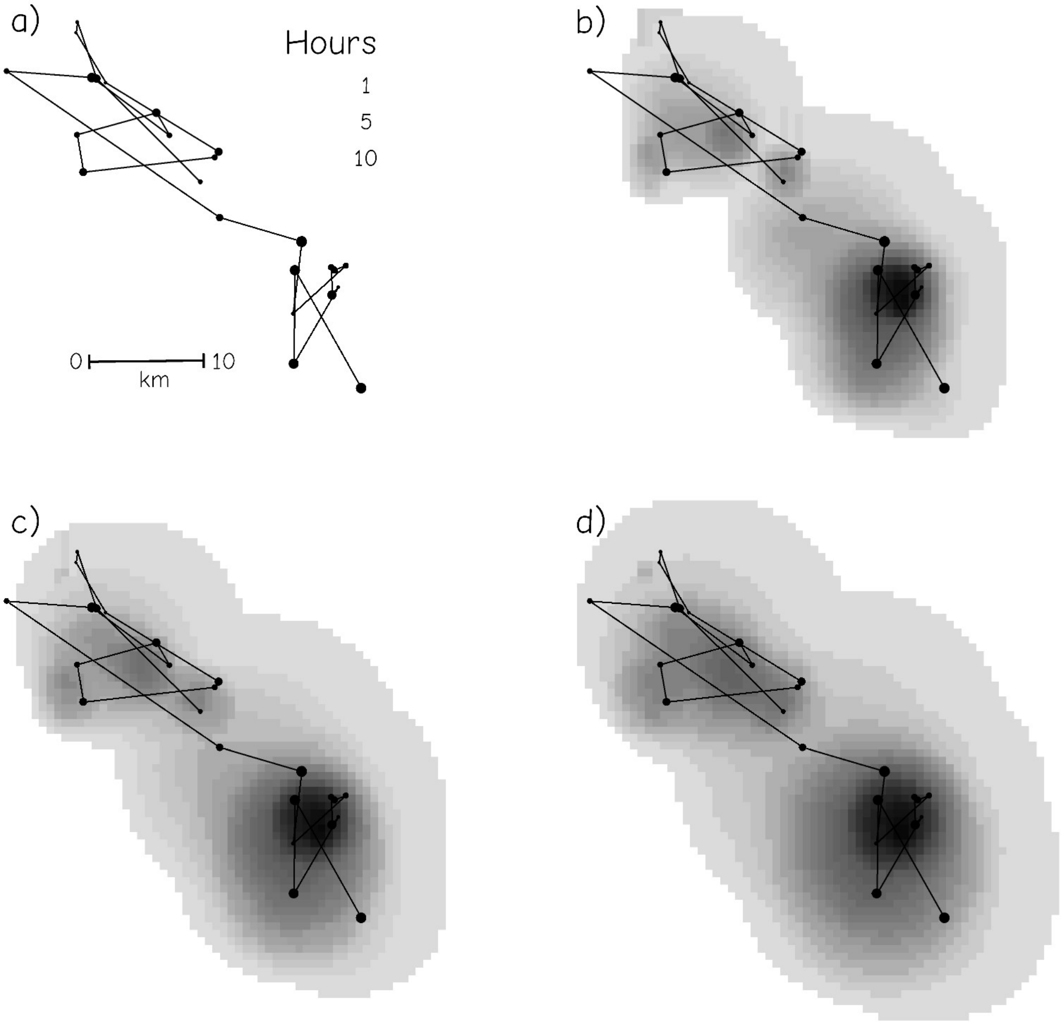


Figure 12-35: An example time-geographic density function for a sequential set of control points. The path (a) shows both the trajectory and interval between observations. Successive figures (b through d) show different maximum velocity values, with velocities of 2 (b), 2.5 (c), and 3 (d) km/hour (courtesy J. Downs).

observations shown by point size. Panels b through d show TGDE calculated using different maximum velocities. Note that the highest densities (darkest shades) show where the control points are clustered and where the distance between observations is short relative to time period between observations, which in turn is dependent on maximum velocities. This implies the net object movement was small between control points, although there is a denser area of likelihood there.

While a maximum velocity may be established from observations or theoretical values, the shape of the distance function often is not. A uniform function may be more defensible if the object is moving at near the maximum speed for most of the duration. However, a linear decay function may make more sense when the sampling interval varies in frequency, and the object is often moving much more slowly than the maximum velocity. TGDE is a developing field, and the interested reader should refer to the papers by J. Downs and colleagues listed at the end of this chapter.

Summary

Interpolation and spatial prediction allow us to estimate values at locations where they have not been measured. These methods are commonly used because our budgets are limited, samples may be lost or found wanting, or because time has passed since data collection. We may also interpolate when converting between data models, for example, when calculating a raster grid from a set of contour lines, or when resampling a raster grid to a finer resolution.

Spatial prediction involves collecting samples at known locations and using rules and equations to assign values at unsampled locations. There are many ways to distribute a sample, including a random selection of sample locations, a systematic pattern, clustering samples, adaptive sampling, or a combination of these. The sampling regime should consider the cost of travel and collecting samples, as well as the nature of the spatial variability of the target feature and the intended use of the interpolated surface.

Sample values are combined with sample locations to estimate or predict values at unsampled locations. There are many spatial prediction methods, but the most common are Thiessen (nearest neighbor) polygon, local averaging (fixed radius), inverse distance weighted, trend surface, and kriging interpolation. Each of these methods has advantages and disadvantages relative to each other, and there is no method that is uniformly best. Each method should be tested for the variables of interest, under conditions in the study area of interest. The best tests involve comparisons of interpolator estimates against withheld sample points.

Measures of core area are commonly identified from spatially distributed observations. This form of prediction identifies regions of high probability for an object or event. Mean center or mean circle are simple measures. A convex hull, defined as the minimum area polygon encompassing all points and with convex exterior angles, is commonly applied. More sophisticated measures include kernel mapping, based on centering scaled distribution functions over each observation, and vertically summing the distribution functions.

Suggested Reading

Anderson, D.J. (1982). The home range: a new nonparametric estimation technique. *Ecology*, 6:103–112.

Angulo-Martínez, M., López-Vicente, M., Vicente-Serrano, S.M., Beguería, S. (2009). Mapping rainfall erosivity at a regional scale: a comparison of interpolation methods in the Ebro Basin (NE Spain). *Hydrology and Earth Systems Science*, 13:1910–1920.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic.

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, 27:93–115.

Anselin, L. (2002). Under the hood: issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 17:247–267.

Anselin, L., Syabri, I., Kho,Y. (2006). GeoDa: An introduction to spatial data analysis. *Geographical Analysis*, 38:5–22.

Ayeni, O.O. (1982). Optimum sampling for digital terrain models. *Photogrammetric Engineering and Remote Sensing*, 48:1687–1694.

Banerjee, S., Carlin, B.P., Gelfand, A.E. (2014). *Hierarchical Modeling and Analysis for Spatial Data (*2nd Ed.). Boca Raton: CRC Press.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 43B:192–225.

Besag, J., Kooperberg, C.L. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82:733–746.

Bowman, A.W., Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford University Press.

Brunsdon, C., Comber, L. (2015). An Introduction to R for Spatial Analysis and Mapping. New York: Sage Publications.

Burgess, T.M., Webster, R. (1984). Optimal sampling strategies for mapping soil types. I. Distribution of boundary spacing. *Journal of Soil Science*, 35:641–654.

Cressie, N. (1991). *Statistics for Spatial Data*. New York: Wiley.

Dale, M.R.T., Fortin, M.J. (2014). *Spatial Analysis, a Guide for Ecologists*. Cambridge: Cambridge University Press.

DeGruijter, J.J., Ter Braak, C.J.F. (1990). Model-free estimation from spatial samples: a reappraisal of classical sampling theory. *Mathematical Geology*, 22:407–415.

Downs, J.A., Horner, M.W. (2009). A characteristic-hull based method for home range estimation. *Transactions in GIS,* 13:527–537.

Downs, J.A, Horner, M.W., Tucker, A.D. (2011). Time-geographic density estimation for home-range analysis. *Annals of GIS*, 17:163–171.

Dubrule, O. (1994). Comparing splines and kriging. *Computers and Geosciences*, 10:327–338.

Fotheringham, A., Brunsdon, C., Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: Sage Publications.

Gelfand, A.E., Diggle, P.J., Fuentes, M., Guttorp, P. (2010). *Handbook of Spatial Statistics*. Boca Raton: CRC Press.

Getis, A., Ord, J.K. (1992). The analysis of spatial association by use of distance statistics, *Geographical Analysis*, 24:189–206.

Goovaerts, P. (1997). *Geostatistics for Natural Resource Evaluation*. New York: Oxford University Press.

Griffith, D.A., Layne, A. (1999). *Casebook for Spatial Statistical Data Analysis*. Oxford: Oxford University Press.

Hutchinson, M.F. (1995). Interpolating mean rainfall with thin plate smoothing splines. *International Journal of Geographical Information Systems*, 9:385–404.

Isaaks, E.H., Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics*. New York: Oxford University Press.

Lam, N.S. (1983). Spatial interpolation methods: a review. *American Cartographer*, 10:129–149.

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, 74:1659–1673.

Mark, D.M. (1987). Recursive algorithm for determination of proximal (Thiessen) polygons in any metric space. *Geographical Analysis*, 19:264–272.

McKillup, S., Dyar, M.D. (2010) *Geostatistics Explained, an Introductory Guide for Earth Scientists*. Cambridge: Cambridge University Press.

Mitasova, H., Hofierka, J. (1993). Interpolation by regularized spline with tension: application to terrain modeling and surface geometry analysis. *Mathematical Geology*, 25:657–669.

Mitchell, A. (1999). *The ESRI Guide to GIS Analysis*. Redlands: ESRI Press.

O'Sullivan, D., Unwin, D.J. (2010). *Geographic Information Analysis,* (2nd Ed.). New York: John Wiley and Sons.

Silverman, B.W. (1986). *Density Estimation*. London: Chapman and Hall.

Varekamp, C., Skidmore, A.K., Burrough, P.A. (1996). Using public domain geostatistical and GIS software for spatial interpolation. *Photogrammetric Engineering and Remote Sensing*, 62:845–854.

Willmott, C.J. (1981). On the validation of models. *Physical Geography*, 2:184–191.

Willmott, C.J., Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30:79–82.

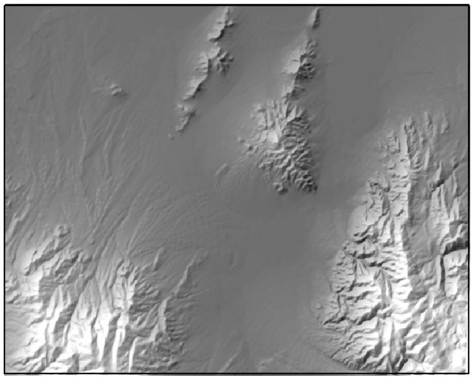
Worton, B.J. (1987). A review of models of home range for animal movement. *Ecological Modellin*g, 38:277–298.

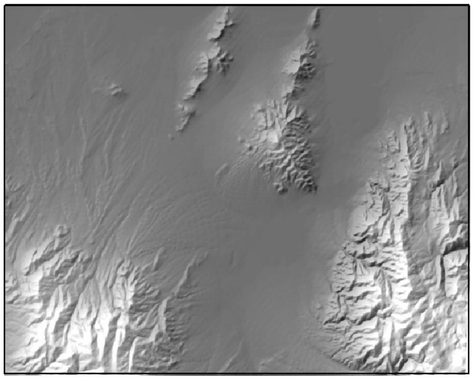
Study Questions

12.1 - Why perform a spatial interpolation?

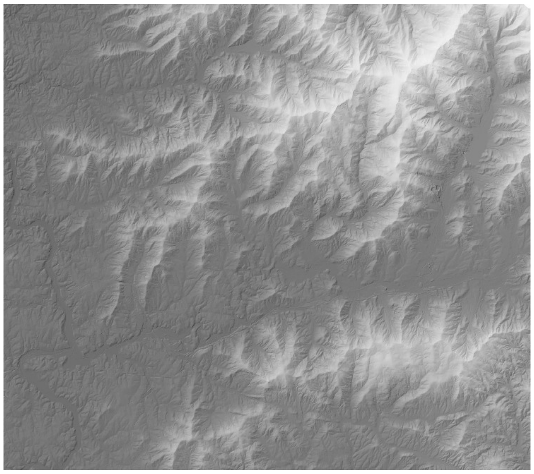
12.2 - Describe four different sampling patterns, and provide the relative advantages or disadvantages of each. Which do you think is used most in practice, and why?

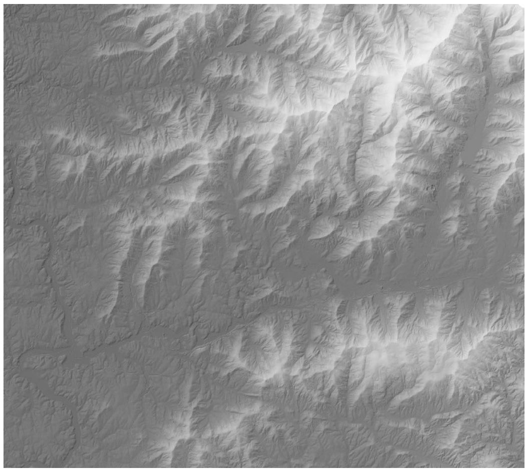
12.3 - Draw a systematic sampling pattern on the area below, left, and an adaptive sampling pattern on the area below, right. Use the same number of sample points, e.g., approximately 50, on both. Which do you think will give a better estimate of terrain locations at unknown points? Why? Would increasing the sample number change which sampling design you would think is best?



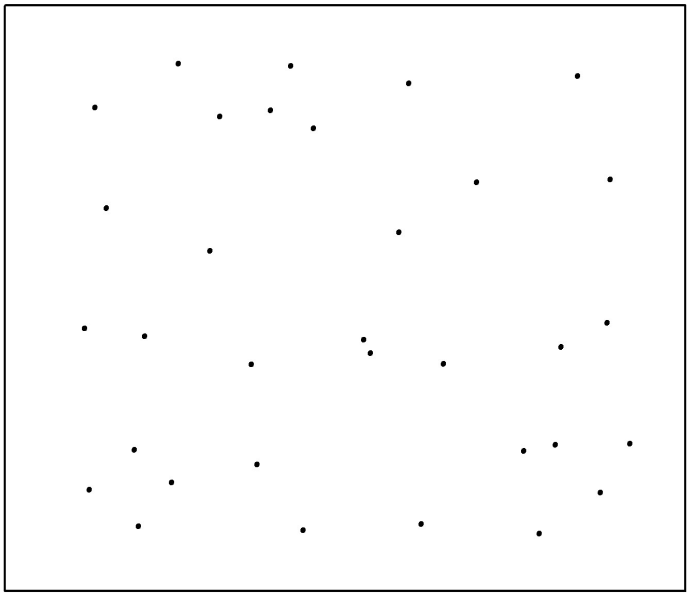


12.4 - Draw a cluster sampling pattern on the area below, left, and an adaptive sampling pattern on the area below, right. Use the same number of sample points, e.g., approximately 50, on both. Which do you think will give a better estimate of terrain locations at unknown points? Why? Would increasing the sample number change which sampling design you would think is best?

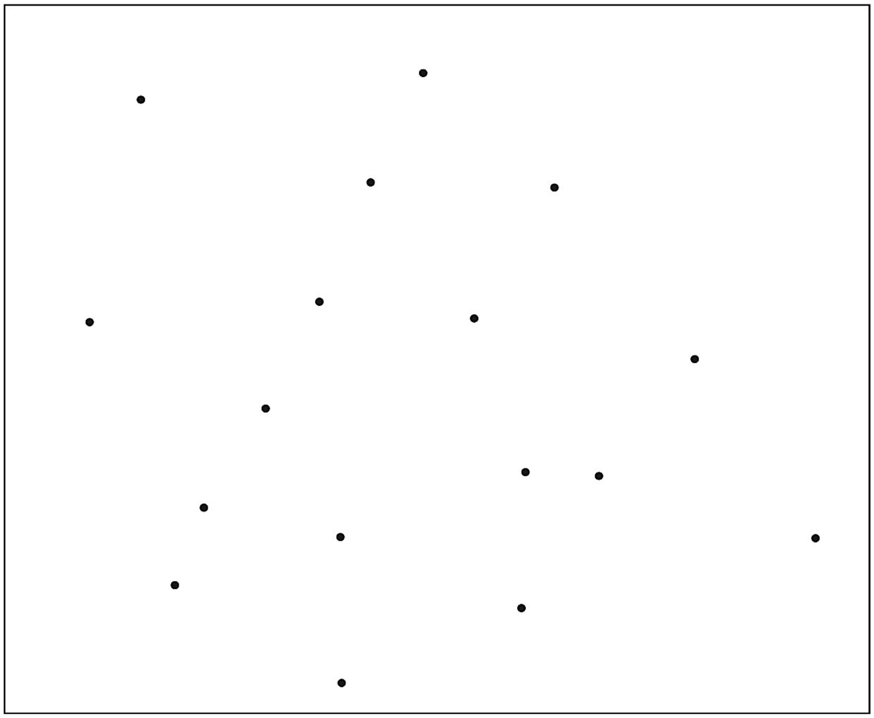




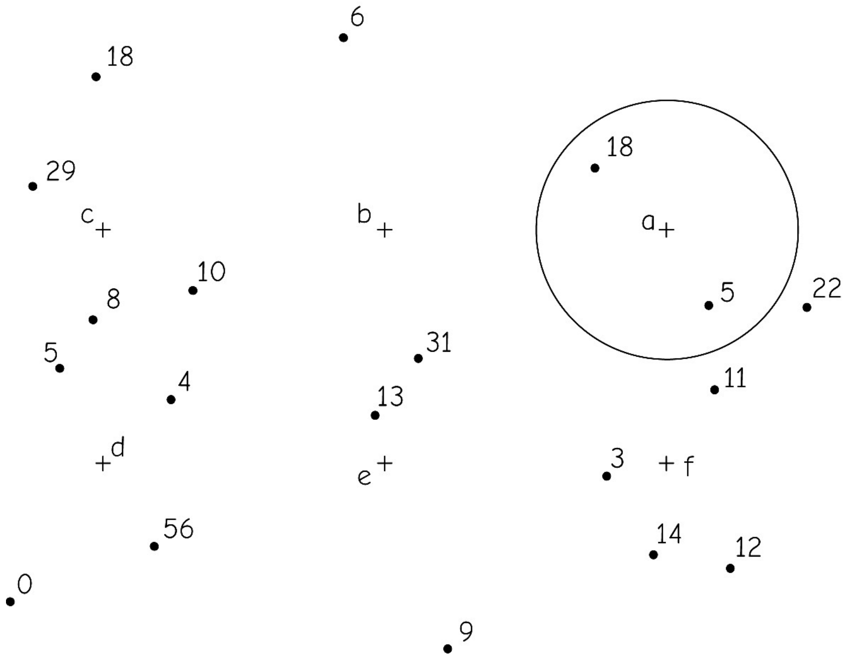
12.5 - Draw the Thiessen polygons (nearest neighbor interpolation) for the set of points below.



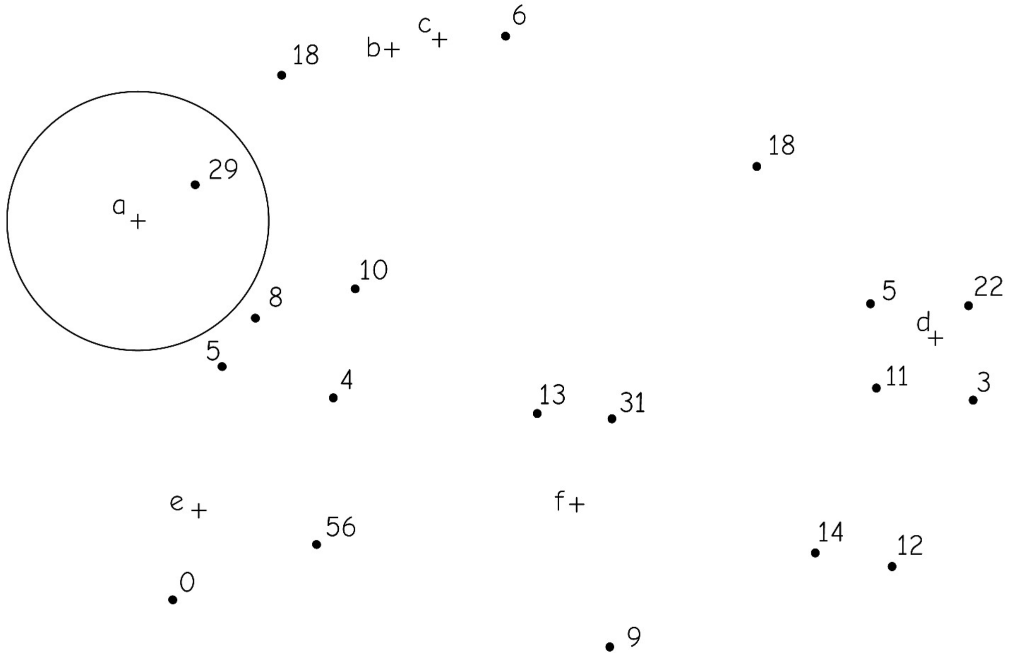
12.6 - Draw the Thiessen polygons (nearest neighbor interpolation) for the set of points below.



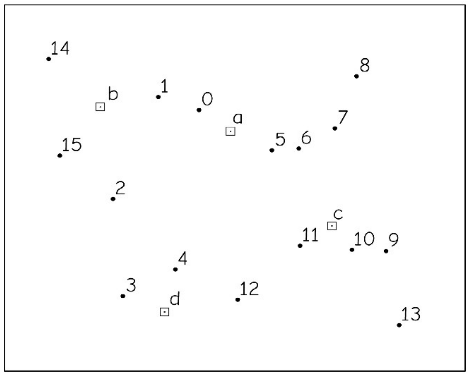
12.7 - Calculate the cell values indicated at the crosses below, using fixed radius sampling size with the shown circle.



12.8 - Calculate the cell values indicated at the crosses below, using fixed radius sampling size with the shown circle.



12.9 - Calculate the Z values for the unknown points listed below, using an inverse distance weighted approach. Use the three nearest known points (use i = 3, n = 1). Known points are shown in map as filled circles and corresponding coordinate and Z values in the table at right.

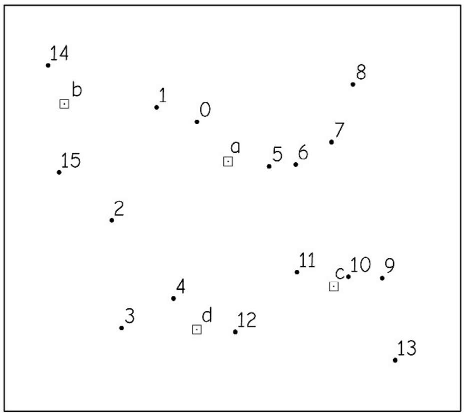


Unknown points:

ID	X	Y	Z
a	155,859.6	4,477,146.0	
b	147,592.6	4,478,884.2	
c	162,515.7	4,469,981.2	
d	151,704.9	4,463,749.2	

ID	X	Y	Z
0	153,951.9	4,478,714.6	2040.6
1	151,280.9	4,479,647.3	1863.0
2	148,228.5	4,472,143.4	1992.1
3	148,906.8	4,464,978.6	2540.1
4	152,383.2	4,466,928.8	2106.3
5	158,827.3	4,475,746.9	2283.2
6	160,607.9	4,475,874.2	1933.5
7	163,024.4	4,477,357.9	1836.4
8	164,465.8	4,481,173.5	1838.3
9	166,416.0	4,468,285.4	2523.9
10	164,169.1	4,468,370.4	2138.8
11	160,692.7	4,468,709.3	1854.2
12	156,537.9	4,464,724.2	1866.9
13	167,306.3	4,462,816.5	2453.8
14	143,946.6	4,482,445.4	1837.9
15	144,709.7	4,475,323.0	1912.8

12.10 - Calculate the Z values for the unknown points listed below, using an inverse distance weighted approach. Use the three nearest known points (use i = 3, n = 1). Known points are shown in map as filled circles and corresponding coordinate and Z values in the table at right.



Unknown points:

ID	X	Y	Z
a	155,859.6	4,477,159.0	
b	147,580.6	4,478,884.2	
c	162,535.7	4,469,960.2	
d	151,714.9	4,463,755.2	

ID	X	Y	Z
0	153,951.9	4,478,714.6	2040.6
1	151,280.9	4,479,647.3	1863.0
2	148,228.5	4,472,143.4	1992.1
3	148,906.8	4,464,978.6	2540.1
4	152,383.2	4,466,928.8	2106.3
5	158,827.3	4,475,746.9	2283.2
6	160,607.9	4,475,874.2	1933.5
7	163,024.4	4,477,357.9	1836.4
8	164,465.8	4,481,173.5	1838.3
9	166,416.0	4,468,285.4	2523.9
10	164,169.1	4,468,370.4	2138.8
11	160,692.7	4,468,709.3	1854.2
12	156,537.9	4,464,724.2	1866.9
13	167,306.3	4,462,816.5	2453.8
14	143,946.6	4,482,445.4	1837.9
15	144,709.7	4,475,323.0	1912.8

12.11 - What is a primary difference between a spline interpolation method and a trend surface interpolation?

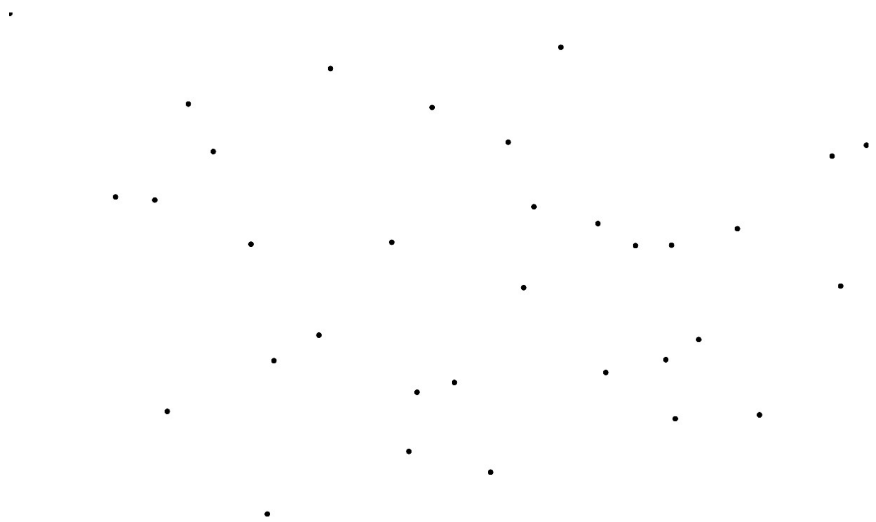
12.12 - What is the primary difference between a trend surface interpolation and a kriged interpolation?

12.13 - Describe the variogram. What does it represent on the X and Y axes, and what are the important regions/points of the plot?

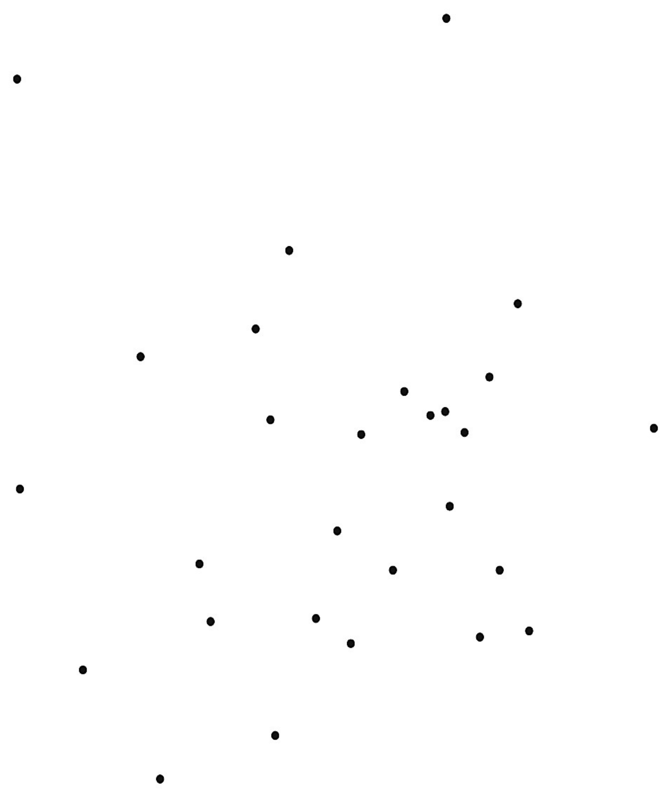
12.14 - Draw the approximate mean center, standard deviation circle, and maximum circle for the following data:

12.15 - What is the convex hull? How is it calculated/determined?

12.16 - Draw the convex hull for the points depicted below:



12.17 - Draw the convex hull for the points depicted below.



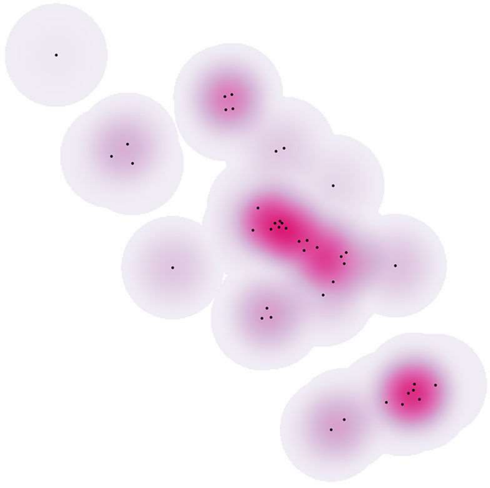
12.18 - Describe/define a kernel density map. Include how the values are based on the samples.

12.19 - Write down and describe at least one equation used to generate a density surface.

12.20 - Which image below illustrates a wider bandwidth? If you used the same data for both plots, give two reasons for your answer, assuming the color scale has the same upper and lower bounds for both surfaces.

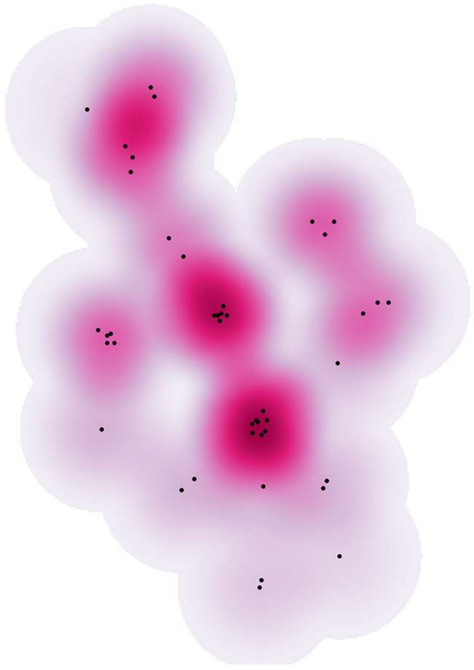
A)

B)

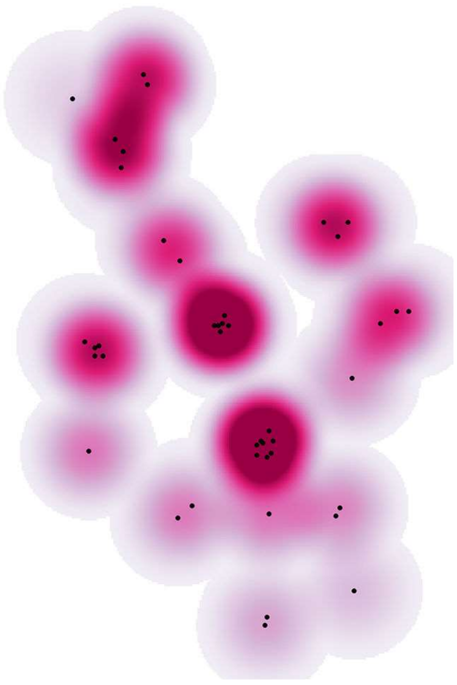


12.21 - Which image below illustrates a larger bandwidth? If you used the same data to generate both plots, give two reasons for your answer, assuming the color scale has the same upper and lower bounds for both surfaces.

A)



B)



13 Spatial Models and Modeling

Introduction

A model is a description of reality. Here, our interest is restricted to computer-based models of spatial phenomena. These models describe the basic properties or processes for a set of spatial features, and help us understand their form and behavior.

Many computer-based models use spatial data, and are developed and run using some combination of GIS, general and specialized computer programming languages, and spatial and non-spatial analytical tools. Spatially explicit models are a primary benefit of GIS technologies, and many spatial models are based on data in a GIS. These models may be run in the GIS, or the spatial data may be prepared in a GIS, and exported to a model that is developed and run outside a GIS.

While there may be as many classes of models as there are modelers, here we split spatial models into three broad and overlapping classes: *cartographic models*, *simple spatial models*, and *spatio-temporal models*. Joseph Berry, an early and well-known developer and proponent of spatial modeling, described cartographic models as automating manual map analysis and processing, while spatial models focus on applying mathematical relationships. Cartographic models are most often applied to rank areas in support of decision making, while simple spatial models often apply sets of equations to predict a specific continuous variable across space. Cartographic model outputs are often nominal (suitable or unsuitable, Figure 13-1) or ordinal (low, medium, or high suitability), while the outputs from simple spatial models are often interval/ratio (e.g., population density, accident frequency, or soil erosion rates).

Cartographic models solve problems via spatial layer combination in overlay, buffers, reclassification, and other spatial operations. These models often employ the concepts of map algebra, described in Chapter 10, but may include a much broader range of operations. *Suitability analyses*, defined here as the classification of land according to their utility for specific uses, are among the most common cartographic models.

Most cartographic models are temporally static because they represent spatial features at a fixed point in time. Data in base layers are mapped for given periods. These data are the basis for spatial operations that may create new data layers. For example, we may be interested in identifying the land that is currently most valuable for agriculture. Costs of production may depend on the slope (steeper is costlier), soil type (some soils require more fertilizer), current land cover (built-up is unsuitable, forests more expensive to clear), or distance to roads or markets. Agricultural production may also depend on soil types, topography (neither flooded nor drought-prone), and the ability to irrigate. Spatial

data on elevation, soil properties, current land use, roads, market location, and irrigation potential may be combined to rank sites by production value. We may use a mathematical relationship for specific calculations of average costs and revenues, for example, agronomists may have developed the relationships between soil types and average corn production in the region, and we may use a cost-per-mile for transport based on local rates. These spatial data are combined in a cartographic model to assign a land value for each parcel in a study region. The model is temporally static in that the values for the spatial variables, such as soil fertility or distance to roads, do not change during the analyses.

Cartographic models are generally not temporally dynamic, even though they may be used to analyze change. For example, we

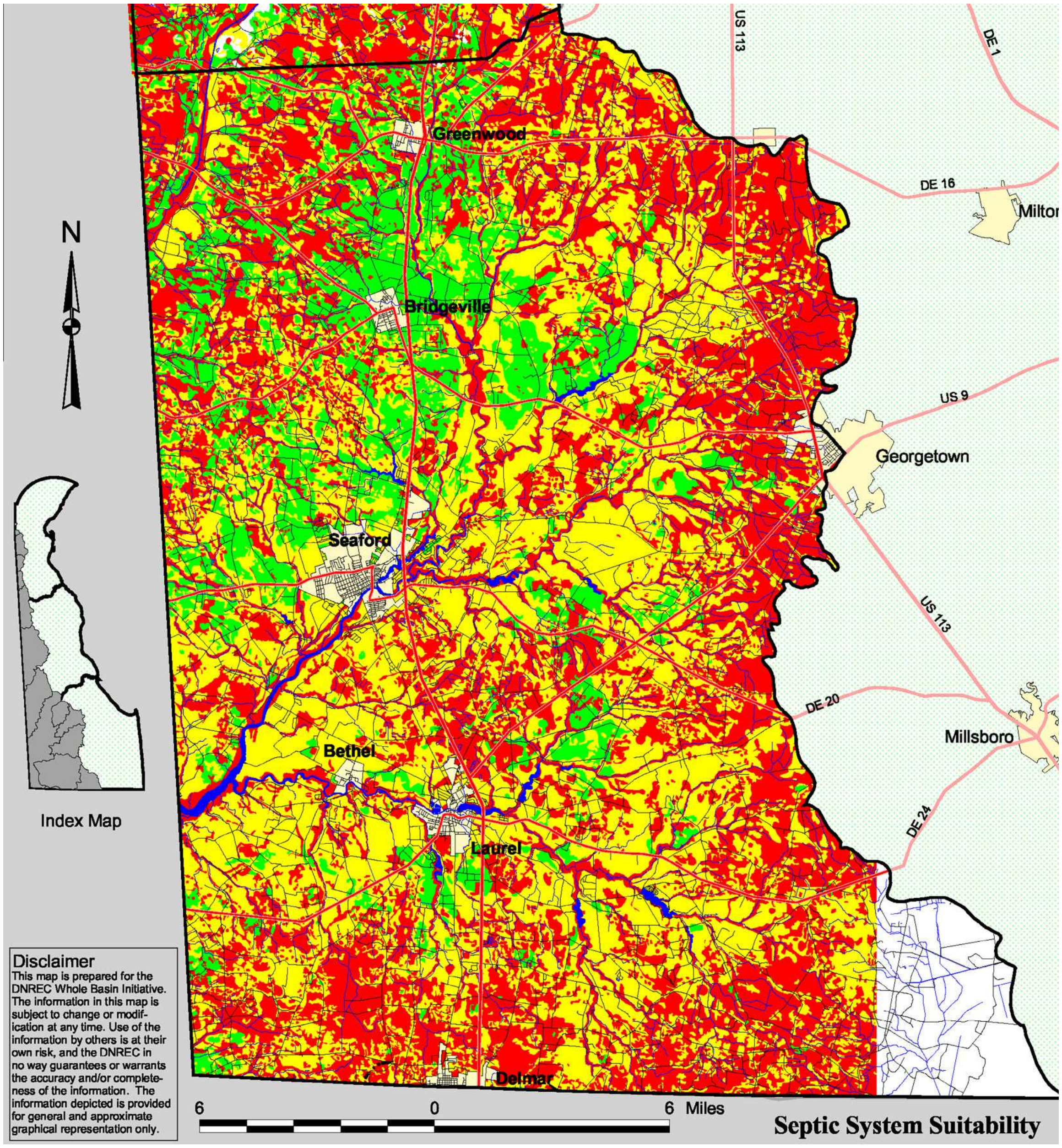


Figure 13-1: An example of a suitability map, produced by combining soils, elevation, wetland, and watercourse data. These analyses are often automated via a cartographic model to produce suitability or other nominal or ordinal rankings over large areas (courtesy state of Delaware).

may wish to analyze vegetation change over a 10-year period, based largely on vegetation maps produced at the start and end of the period. Each data layer represents the vegetation boundaries at a fixed point in time. The model is static in that the polygon boundaries for a given layer do not change. There may be two vegetation data layers, each corresponding to a different point in time, and the vegetation boundaries are mapped as found at each time interval. Our cartographic model includes a temporal component in that it compares vegetation change through time, but the cartographic model does not generate new boundaries of polygons or any other characteristics of spatial features. Boundaries may be a composite of those lines that exist in the input data layers, but new lines at new coordinate locations are not generated. Most spatial modeling or models conducted in the framework of GIS have been cartographic models that are temporally static in this manner.

Simple spatial models typically apply a set of equations to spatially resolved variables (Figure 13-2). They often rest on equations developed from data at a set of observations at points or sub areas, and then applied across broader geographic areas.

An example may help understand simple spatial models. William Cooke and colleagues reported on a model of West Nile virus infection among birds, and the risk of transmission to humans. West Nile virus is a sometimes fatal disease, and varies in prevalence through space and time. Cooke and his associates compiled data on the frequency of bird and human infections within each zip code in Mississippi over several years. Human and bird cases were clustered, with outbreaks concentrated in rural areas. Road density was used as a surrogate for rural/urban landuse.

Spatial variables related to mosquito habitat quality were compiled statewide, including stream density, vegetation type, temperature, and precipitation surplus. These were combined with virus infection frequency at specific locations to fit a predictive statistical model. Mapped spatial variables were then applied in the model to predict outbreak risk across the state.

Simple spatial models are common, with hundreds of examples found across a range of disciplines. They typically include a model derived through sampling and a statistical fitting process, a model that is subsequently applied across space to estimate important events, densities, or other characteristics.

Spatio-temporal models are dynamic in both space and time. They differ from cartographic or predictive spatial models in that time passes explicitly within the running of the model, and changes in time-driven processes within the model cause changes in

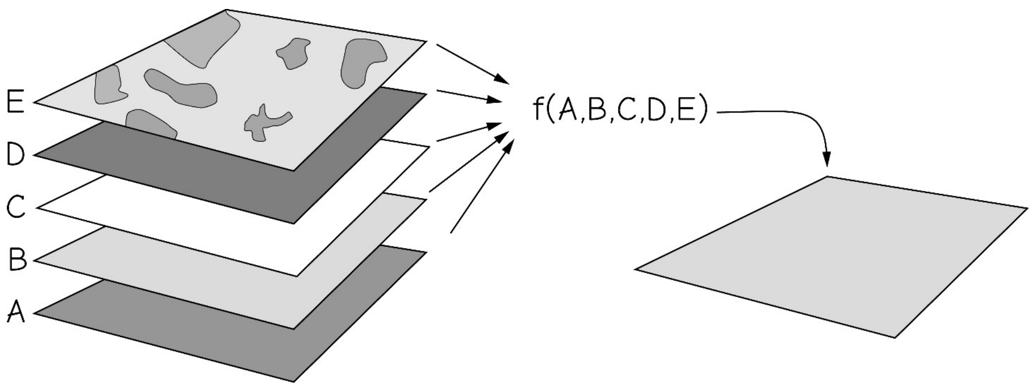


Figure 13-2: An example of a simple spatial model, where a function is applied to spatial inputs to estimate an important spatial output.

spatial variables. Spatio-temporal models often attempt to explicitly represent processes within the model.

The dispersion of oil after a spill is an example of a process that might be analyzed via a spatio-temporal model. Currents, winds, wave action, and the physics of oil separation and evaporation on exposure to air might be combined in a model to predict the changing location of an oil slick. The actions of objects as they move across an environment may also be represented in a spatio-temporal model.

Spatio-temporal models include time-driven processes within the framework of the model. These processes are typically quite detailed and include substantial computer code to represent important subprocesses. Our oil evaporation example demonstrates the subprocesses represented in a dynamic spatial model. Oil evaporation rates depend on many factors, including oil viscosity, component oil fractions, wind speed, temperature, wave height and action, and sunlight intensity. These processes may be modeled by suitable functions applied to spatially defined patches of oil. The submodel may estimate evaporation of various components of the oil in the patch, and update the characteristics of oil in that patch. Oil chemistry and viscosity may change due to more rapid evaporation of lighter components, in turn affecting future evaporation calculations. Spatial features may change through time due to the represented dynamic process; for example, the boundary defining an oil spill may vary as the model progresses.

Spatio-temporal models are typically more limited than other modeling approaches in the range and number of spatial themes analyzed, but they provide a more mechanistic representation of dynamic processes. Substantial effort goes into developing submodels of important processes. Model components and structures focus on one or a few key output spatial variables, and input data themes are included only as they are needed by these subprocess models. These temporally dynamic models explicitly calculate the changes in the output spatial variables through time. Feature boundaries, point feature locations, and attribute variables that reflect the spatial and aspatial characteristics of key output variables may change within the model run, typically multiple times, and with an explicit temporal frequency.

Simple spatial models and spatial statistical analyses are often used as precursors to spatio-temporal models. By uncovering key processes or rates, they can guide further analysis. For example, in our oil spill example, the specific relationship between wave height or frequency and oil separation may be represented by an equation, but the specific parameters that define the shape of the relationship may be estimated via a statistical process. Experiments or observations on separation rates at various wave heights may be collected, and the specific model parameters estimated. These may then be included as a component of the larger spatio-temporal model.

Cartographic Modeling

A *cartographic model* provides information through a combination of spatial data sets, functions, and operations. These functions and operations often include reclassification, overlay, interpolation, terrain analyses, buffering, and other functions. Multiple data layers are combined via these operations, and the information is typically in the form of a spatial data layer. Map algebra, described in Chapter 10, is often used to specify cartographic models for raster data sets.

Suitability analyses are perhaps the most common examples of cartographic models. These analyses rank land according to its utility for various purposes. Suitability analyses often involve the overlay, weighting, and rating of multiple data layers to categorize lands into various classes. Relevant data layers are combined and the resultant polygons are classified based on the combination of attributes. Figure 13-3 illustrates a simplistic cartographic model for the identification of potential park sites. Suitable sites are those that are near lakes, near roads, and not in wetlands. The model uses three input data layers, containing lakes, roads, and hydric status for a common study area. Spatial operations are applied to the spatial data layers, including reclassification, buffering, and overlay. These result in a suitability layer. This suitability layer can then be used to narrow sites for further evaluation, identify owners, or otherwise aid in park site selection.

Cartographic models have been used for a variety of applications. These include landuse planning, transportation route and corridor studies, the design and development of water distribution systems, modeling the spread of human disease or introduced plant and animal species, building and business site selection, pollution response planning, and endangered species preservation. Cartographic models are so extensively used because they provide information useful to managers, the public, and policy makers, and help guide decisions requiring the consideration of spatial location across multiple themes.

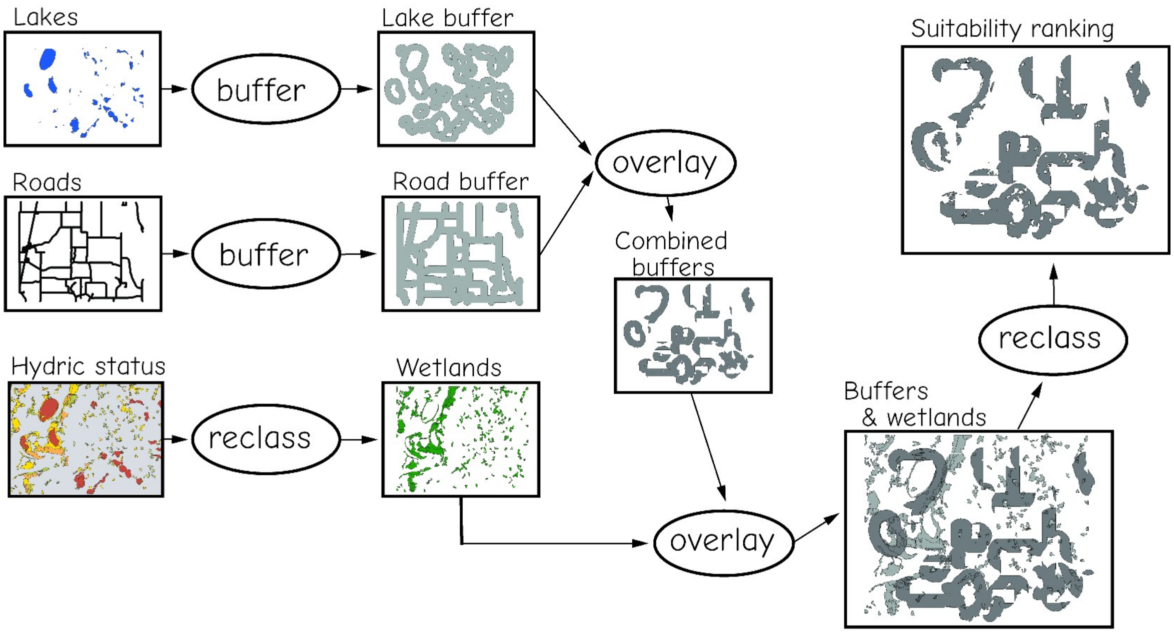


Figure 13-3: An example of a cartographic model. The model identifies suitable park sites based on the proximity to roads and lakes, and the absence of wetlands.

Cartographic models are often succinctly represented by *flowcharts*. A flowchart is a graphic representation of the spatial data, operations, and their sequence of use in a cartographic model. Figure 13-4 illustrates a flowchart of the cartographic model illustrated in Figure 13-3. Suitable sites are sought that are near roads, near lakes, and not in wetlands. Data layers are represented by rectangles, operations by ellipses, and the sequence of operations by arrows. Operations are listed in each ellipse. Flowcharts are often required by an agency or organization to document a completed spatial analysis. Because a consistent set of symbols aids in the effective communication of the cartographic model, a standard set of symbols and flowcharting methods may help in understanding the data and operations in an analysis.

Flowcharts are useful during the development and application of a cartographic model. Flowcharts aid in the conceptualization and comparison of various competing approaches and may aid in the selection of the final model. A flowchart is often an efficient framework for documenting a cartographic model. File locations, work dates, and intermediate observations can be noted with reference to the flowchart, or directly onto a copy of the flowchart.

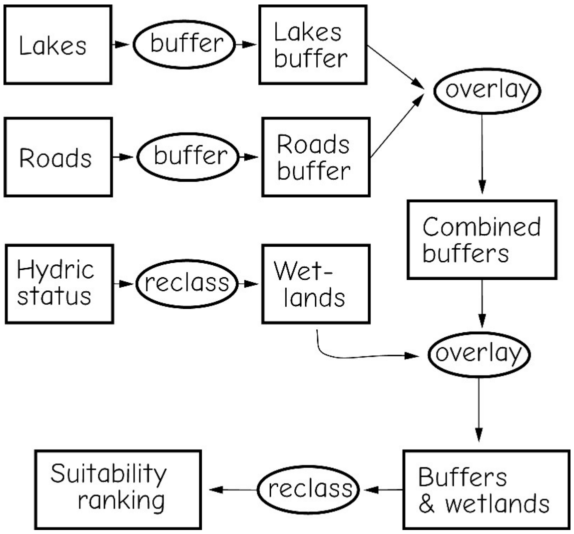


Figure 13-4: A flowchart of the cartographic model in Figure 13-3. The flowchart is a shorthand method of representing a sequence of spatial operations.

Cartographic modeling often produces a large number of intermediate or temporary data layers that are not required in the final output. Our example in Figure 13-4 illustrates this. The needed information is contained entirely within the suitability ranking data layer. Five other data layers were produced within the illustrated cartographic model. Buffered, recoded, and overlay layers were necessary intermediate steps, but in this analysis their utility was temporary. This proliferation of data layers is common in cartographic modeling, and it can cause problems as the new layers and other files accumulate in the computer workspace. Frequent removal of unneeded files is helpful.

Much of the power of cartographic modeling comes from the flexibility of spatial analysis functions. Spatial functions and operations are a set of tools that may be mixed and matched in cartographic models. Overlay, proximity, reclassification, and most other spatial analysis tools are quite general. These tools may be combined in an astoundingly large number of ways, in selection and order of application. These variations will result in different output data layers, even when using the same input data layers. With a small set of tools and data layers, we can create a huge number of cartographic models. Designing the best cartographic model to solve a problem — the selection of the appropriate spatial tools and the specification of their sequence — is perhaps the most important and often the most difficult process in cartographic modeling.

Designing a Cartographic Model

Most cartographic models are based on a set of criteria. Unfortunately, these criteria are often initially specified in qualitative terms, such as “the slopes must not be steep.” A substantial amount of interpretation may be required in translating the criteria in a suitability analysis into a specific sequence of spatial operations. In our present example, we must quantify what is meant by “too steep.” General or qualitative

criteria may be provided and these must be converted to specific, quantitative measures. The conversion from a qualitative to quantitative specification is often an iterative process, with repeated interaction between the analyst developing and applying the cartographic model and the manager or decision-maker who will act on the resultant information.

We will use a home-site selection exercise to demonstrate this process. The problem consists of ranking sites by suitability for home construction. The area to be analyzed has steep terrain and is in a seasonally cold climate. There are four criteria:

a) Slopes should not be too steep. Steep slopes may substantially increase costs or may preclude construction.

b) A southern aspect is preferred, to enhance solar warming.

c) Soils suitable for on-site septic systems are required. There is a range of soil types in the study area, with a range of suitabilities for septic system installation.

d) Sites should be far enough from a main road to offer some privacy, but not so far as to be isolated.

These criteria must be converted to more specific restrictions prior to the development and application in a cartographic model. The decision-maker must specify what sort of classification is required. Is a simple binary classification needed, with suitable and unsuitable classes, or is a broader range of classes needed? If a range of classes is specified, is an ordinal ranking acceptable, or is an interval/ratio scale preferred? These questions are typically answered via discussions between the analyst and the decision-makers. Each criterion can then be defined once the type and measurement scale of the results are specified. It may be fairly simple to establish the local slope limit that prohibits construction. For example, conversations with local building experts may identify 30 degrees as a threshold beyond which construction is infeasible. Further work is required to quantify how slopes affect construction costs. Similar refinements must be made for each criterion. We must quantify the range and any relative preferences for southern aspects, relative soil suitabilities, what defines a main road, and what constitutes short and long distances.

A second key consideration involves the availability and quality of data. Do the required data layers exist for the study area? Are the spatial accuracies, spatial resolution, and attributes appropriate for the intended analysis? How will map generalizations affect the analysis; for example, will inclusions of different soil types in a soil polygon lead to inappropriate results? Is the minimum mapping unit appropriate? If not, then the requisite data must be obtained or developed, or the goals and cartographic model modified.

Weightings and Rankings

While some cartographic models are simple and restrictive, many more cartographic models require the combination of criteria that vary across a range of values, and require an explicit ranking of the relative importance of different classes or types of criteria. A simple, restrictive example might require us to identify parcels greater than a certain size and within a certain distance of water. We may clearly identify areas that meet these desired conditions.

A much more common class of problems requires us to integrate multiple criteria that are qualitatively different. For example, site suitability for hazardous waste storage depends on a number of factors, including distance to population centers, transportation, geology, and aquifer depth and type. We must rate sites across a range of values for all of these variables. Once criteria are precisely defined, we must obtain appropriate data, develop a flowchart or plan for our analysis, and address the more difficult problem of assigning rankings within each criteria, and assigning the relative weightings among criteria. Note that in the following discussions we use the word “rankings” when describing the assignment of relative values within the same layer, such as how

we rank a sandy soil vs. a silty soil in a soils layer. We use the word "weightings" when assigning the relative values of different layers, for example, how we weight the values in an elevation layer vs. the values in a landuse layer.

Rankings Within Criteria

Each criterion in our cartographic model is usually expressed by a data layer, or "criterion layer." Each criterion layer is a spatial representation of some constraint or selection condition; for example, the criterion we build outside a floodplain may consist of a set of numbers in a layer identifying floodplain locations. Floodplain sites may be assigned a value of 0, and upland sites a value of 1.

Before we can assign a value to any site, we must first obtain floodplain maps and interpret the codes in the maps to delineate the most flood-prone areas. Floodplain maps may exist, or we may have to generate them from other sources. This allows us to rank areas based on the likelihood of flooding.

We must explicitly formalize our ranking for each layer used to represent a criterion. One early decision is whether ranks should be discrete or continuous (Figure 13-5). Rankings are discrete when input data are interpreted such that the criterion data layer is a map of discrete values. Soils are either good or bad for construction, slopes either too steep or acceptable, and the final map defines two or a few discrete classes; for example, sites are categorized as either suitable or unsuitable. Ranks are continuous when they vary along a scale; for example, soils may be rated from 1 to 100 for construction suitability.

Figure 13-5, top right, shows the assignment of discrete ranking of land productivity based on values in a soil layer. The source layer in the top left of the figure is analyzed. If the expected production for a given soil

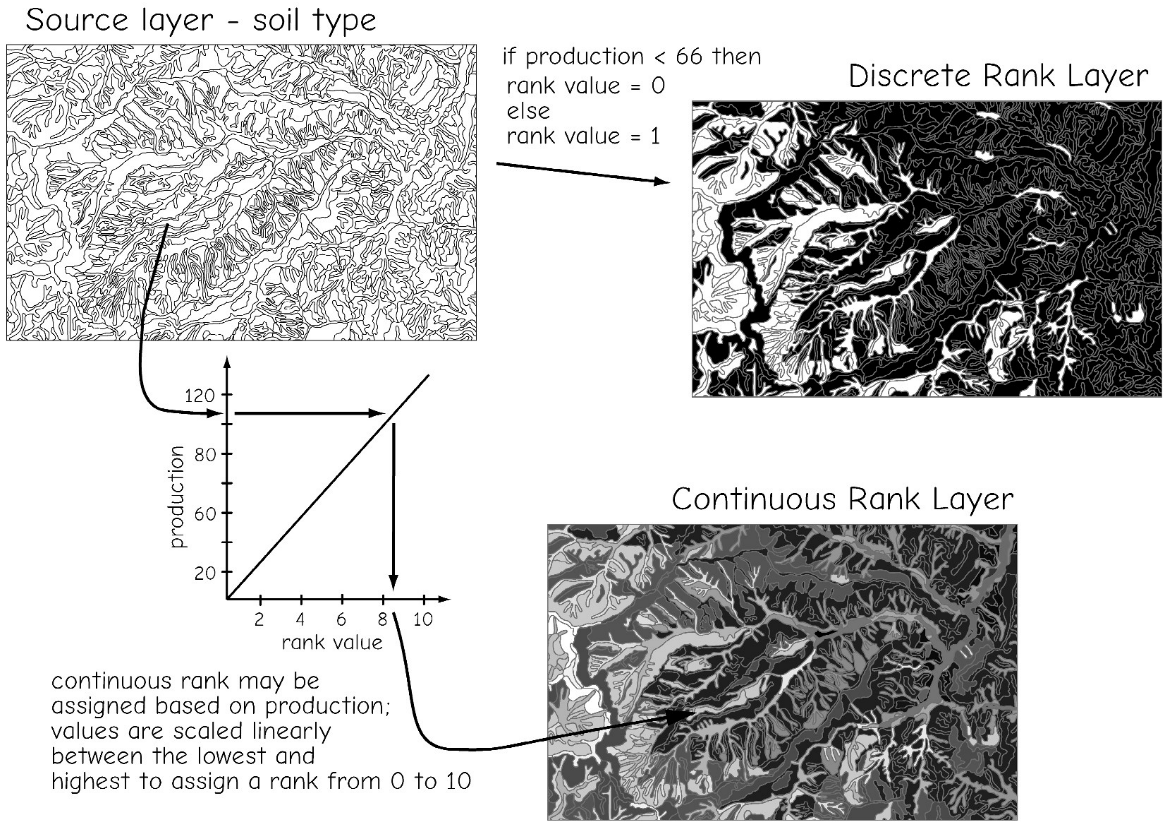


Figure 13-5: Rankings within a layer are often discrete (top right) or continuous (bottom).

polygon is less than 66, then the output ranking is set to 0. If the production is greater than or equal to 66, then the output rank is set to 1. A range of input values has been placed into two discrete classes, illustrated as the discrete rank layer in the top-right part of Figure 13-5.

Discrete rankings are most often used when there are clear, discrete classes to be represented in criteria. A disease may be present or absent, a country an ally or enemy, or a block inventoried or not. The values to be represented are discrete categories.

In contrast, we may apply criteria as continuous rankings within a cartographic model. These continuous rankings provide a range of values to characterize a suitability or restriction, and they result in a set of incrementally varying ranks. Ranks (or scores) typically range over a real or large integer interval, for example, from 0 to 1 or 0 to 1,000. Highest suitability is usually assigned to the highest rank, and lowest to the bottom.

The bottom right of Figure 13-5 shows a continuous ranking over a range of 0 to 10. A high value of 10 is specified for the most productive soils, and a low value of 0 for the least productive. We may use production data gathered over a set of soil types, and a map of soil types to assign the relative value of each soil. We could scale production from the lowest to the highest observed over the range of 0 to 10, and in so doing create a layer that represents a soil productivity criteria.

We are not constrained to linear or always increasing or decreasing relationships between our input layers and our criteria layers. There may be complex relationships between an input value and our output ranking scores or values. Any curve or relationship we can create with a combination of mathematical and logical functions may be represented, to reflect increasing, decreasing, or complex relationships.

We should have some justification for adopting a specific curve when establishing relative ranks within a layer. For example, we may wish to represent the mercury hazard based on methyl mercury concentrations in water supplies across a state. There may be a broad range of low mercury concentrations for which there are no or few negative health impacts. However, as a threshold concentration is reached, there may be a rapid upturn into a very steep curve, where the risk of severe damage is great (Figure 13-6). The shapes of these curves should be established through sets of epidemiological studies, in which mercury concentration in human blood or tissue was related to drinking water, and health impacts were recorded for thousands of people at various levels of mercury exposure.

Figure 13-7 illustrates two examples of continuous criteria scores. Figure 13-7a shows the representation of a complex road criterion for a cartographic model. This criterion specifies that desirable sites are greater than 300 but less than 2,000 m from a road. The top left graphic of Figure 13-7a shows the original roads layer. Following the arrows counter clockwise, you find the distance layer, a raster with the distance from

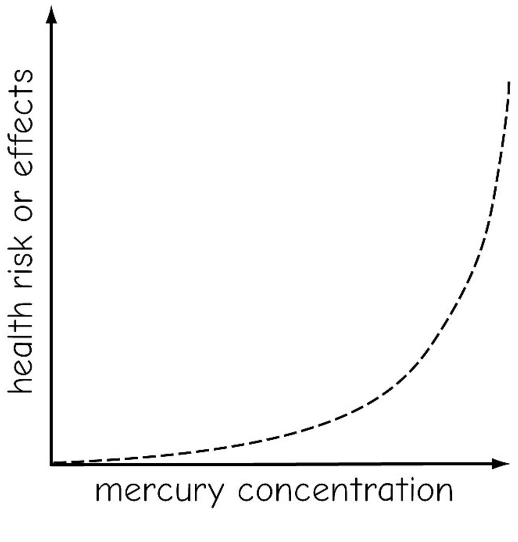


Figure 13-6: The ranking within a criterion layer should be based on a defensible relationship, whose shape has been established through sufficient study or experience. Here, risk for mercury exposure via concentration in drinking water has been related to negative health impacts. Suitability or hazard rankings in cartographic models should be well supported by measurements.

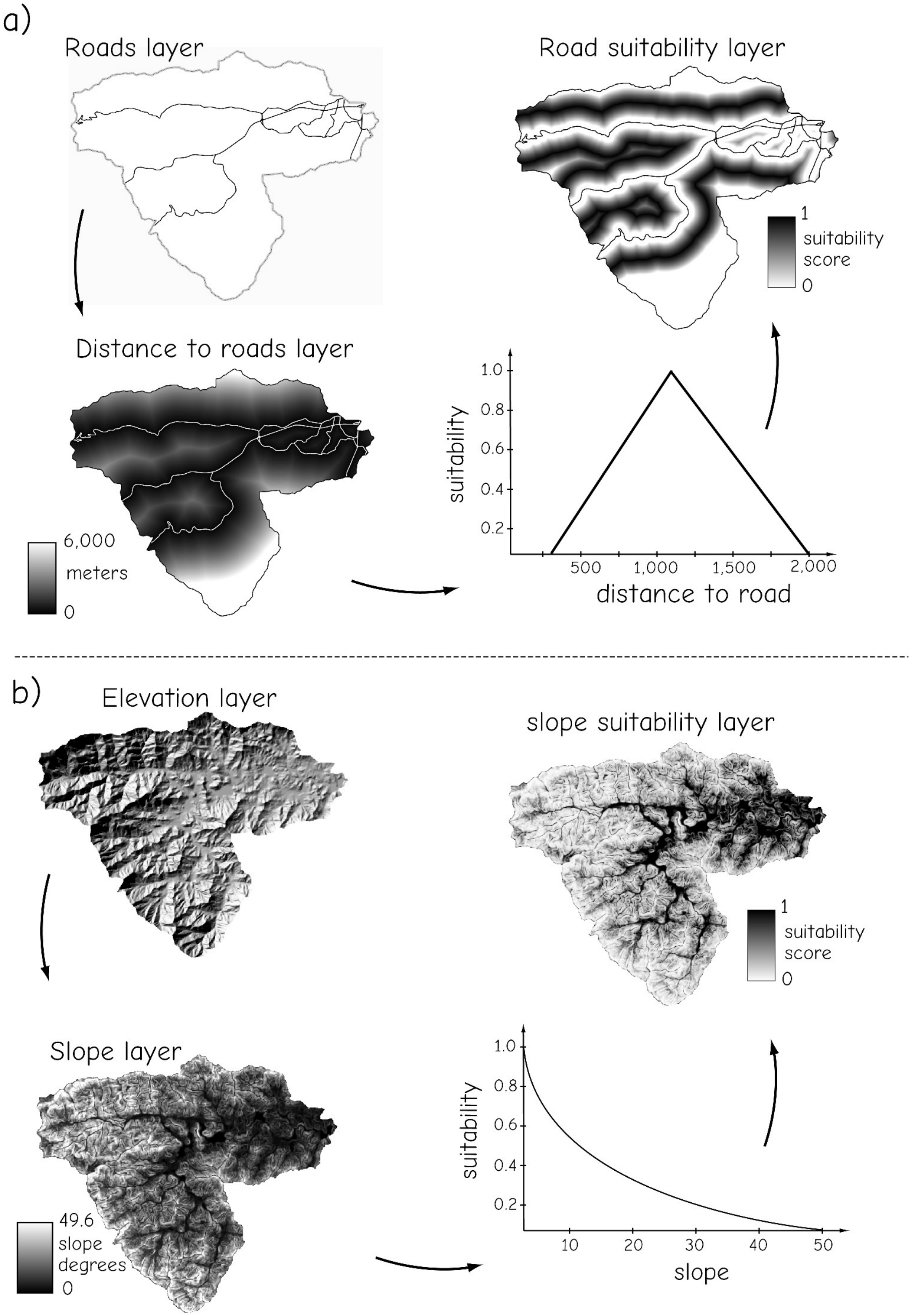


Figure 13-7: Examples of two continuous score layers to represent criteria in a cartographic model. Panel a, top, shows the dual distance to road criteria applied using a suitability function. This results in a continuous range of suitability scores between 0 and 1. In panel b, a different suitability function is applied to the slope layer, resulting in a continuous suitability score for the slope criterion.

the nearest road recorded in each cell value. In this example the distances range from 0 to 6,000 m. The graphic in the lower right of Figure 13-7a shows a suitability assignment function. Distance values are recorded along the horizontal axis, and are used to assign suitability for building, shown on the vertical axis. This function assigns suitability scores of 0 for distances less than 300 m. Suitabilities increase and distance increases, in a linear fashion, to a score of 1 at a distance of 1,150 m half way between 300 and 2,000. Scores then decline linearly to a value of 0 at 2,000 m, and remain 0 for all distances greater than 2,000 m.

Figure 13-7b illustrates a continuous ranking of suitability scores, in this instance for slope. Slopes are calculated from the elevation layer (Figure 13-7b, left), ranging from 0 to 49.6 degrees for this data set. Slope values are transformed to continuous slope suitability values using a smoothly decaying function (lower right, Figure 13-7b). These values are assigned to each cell location in an output slope suitability data layer (top right, Figure 13-7b).

Note that these continuous rankings may be combined, often through a weighted addition process, to generate a combined suitability score. The various suitability layers sum vertically to give a total composite score for each cell. This score may be used to rank areas on relative suitability. Discrete and continuous suitability layers may be combined using a mix of Boolean and addition operations to provide a final ranking. This combination often requires that we define the relative importance of each criteria layer, a process known as weighting among criteria.

Weighting Among Criteria

Distinct criteria must be combined in many spatial analyses, usually in some overlay or addition process (Figure 13-8). We must choose how to weight one layer relative to another. How important is slope relative to aspect? Will an optimum aspect offset a moderately steep site? How important is isolation relative to other factors? Because the criteria will be combined in a suitability data layer, the relative weightings given each criterion will influence the results. Different

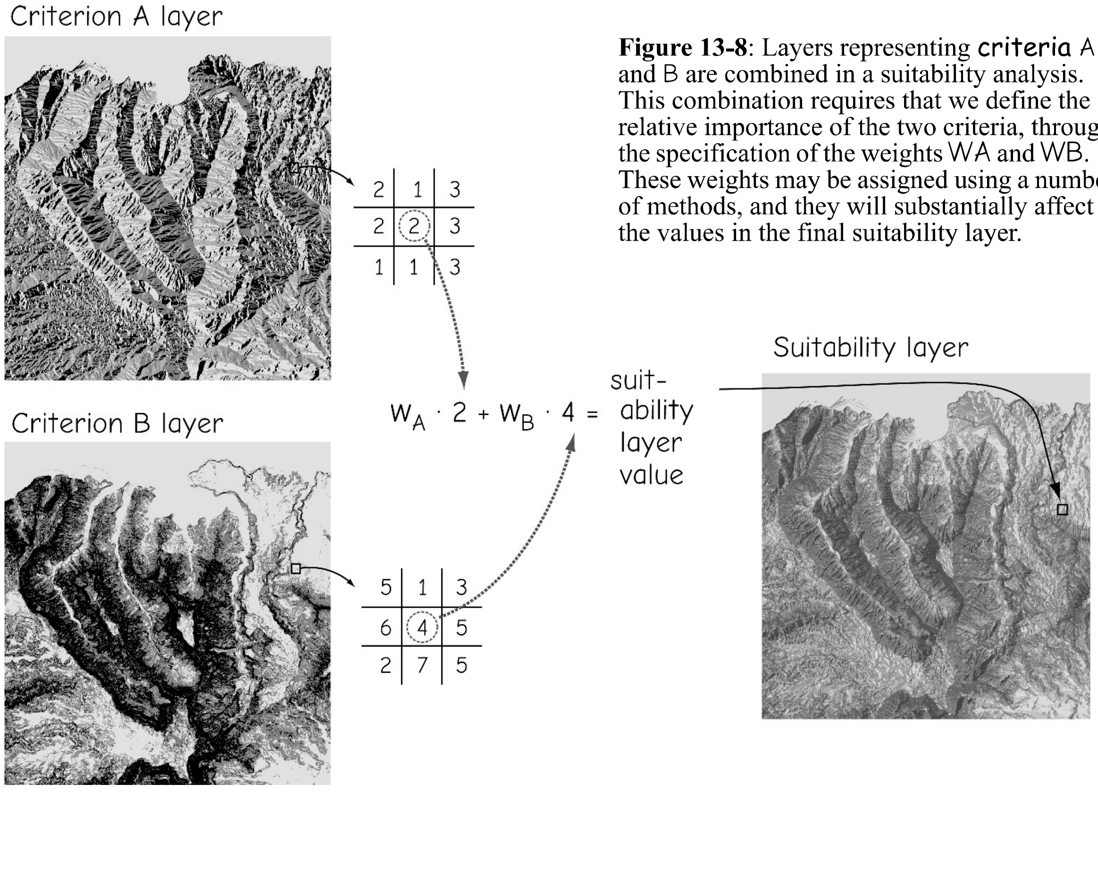


Figure 13-8: Layers representing criteria A and B are combined in a suitability analysis. This combination requires that we define the relative importance of the two criteria, through the specification of the weights WA and WB. These weights may be assigned using a number of methods, and they will substantially affect the values in the final suitability layer.

relative weights are likely to result in different suitability rankings. It is often difficult to assign these relative weights in an objective fashion, particularly when suitability depends on nonquantifiable measures.

The assignment of relative weightings is easiest when the importance of the various criteria may be expressed on a common scale. In our example, we may be able to assign monetary costs to increasing slope. Soil types may be categorized based on their septic capacity. Different septic systems may be required for different soil types, either through larger drain fields or the specification of mound vs. field systems. Costs could be estimated based on the variable requirements set by soil type. Nuisance cost for noise and distance cost in lost time or travel might also be quantified monetarily. Reducing all criteria to a common scale removes differential weighting among criteria.

There are many instances where a common measurement scale is not possible. Many rankings are based on variables that are difficult to quantify. Personal values may define the distances from a road that constitute “isolated” versus “private,” or what is the relative importance of slope vs. construction cost. Expert opinion, group interviews, or stakeholder meetings may be used to rank when there are multiple or competing parties. The scales for these variables are inherently different, and there is no clear way to translate them to a common scale.

One method of assigning weights is based on their “importance ranking.” The factors (criteria) used to decide the quality of a site may be ranked in their importance, from most important to least. We may then calculate the relative weights according to:

$$w_i = \frac{n - r_i + 1}{\sum_{k=1}^{n} (n - r_k + 1)} \qquad (13.1)$$

where w_i is the weighting for criterion i, n is the number of criteria, and k is a counter for summing across all criteria.

Suppose we wish to rank sites for store placement based on four factors: distance to nearest competitor, distance to nearest major road, parking density, and parcel cost. Figure 13-9 shows an example calculation of criteria weights based on importance ranking. Each criterion is listed in the leftmost column. Ranks are assigned to criteria by the planner, client, decision-maker, or interested group. The numerator of equation 13.1 is calculated for each criterion, giving the most important criterion the highest value and the least important the lowest value. The denominator is calculated by summation, and then the individual weights calculated, as shown in the right most column. These weights may then be used to combine the data from the various criteria.

Note that there are several assumptions in this example. First, we assume that the values in each layer associated with each criterion have appropriate ranges, or at least are on similar scales. In Figure 13-8, the values for criterion layer A and criterion layer B vary over an approximately equal range. If one layer had a range from 1,000 to 5,000 and the other had values of 1 to 5, then this would affect the combination and final suitability ranking.

Second, we may be implicitly assuming that the scales are approximately linear in our ranking within and across the criteria. We often combine the values within a criterion layer using an arithmetic operation, for example, by summing values with weights (Figure 13-9). The relative weights among and within each layer are mixed, which is often a logical course of action under an assumption of linearity. Strongly nonlinear relationships in the ratings and weightings scales often lead to counterintuitive and unwanted suitabilities.

There are many other methods for defining the values for each criterion layer and the relative weightings among layers. These include methods that attempt to ensure con-

Criterion	Rank	Numerator $(n - r_i + 1)$	Weight $(n - r_i + 1) / \sum_{k=1}^{n}(n - r_k + 1)$
distance to nearest competitor	2	4-2+1 = 3	3/10 = 0.3
distance to major road	3	4-3+1 = 2	2/10 = 0.2
parking density	4	4-4+1 = 1	1/10 = 0.1
parcel cost	1	4-1+1 = 4	4/10 = 0.4
		$\sum_{k=1}^{n}(n - r_k + 1) = 10$	

Figure 13-9: An example of one method for calculating relative weights for each of four criteria according to equation (13.1).

sistency among weights, but they are beyond the scope of this introductory text. You may find more detailed descriptions in the excellent book by Malczewski (1999) listed in the suggested readings section at the end of this chapter.

Cartographic Models: A Detailed Example

Here we provide a detailed description of the steps involved in specifying and applying a cartographic model. We use a refinement of the general criteria for homesite selection described in the previous section. These general criteria are listed on the left side and the refined criteria are shown on the right side of Table 13-1. The refined criteria may have been defined after further discussion with the decision-makers, local area experts, and a review of available data and methods.

Note that we adopt the simplest weighting and ranking scheme in applying the criteria in Table 13-1. All criteria are equally weighted, and all criteria are binary — land is categorized as unsuitable or suitable based on each criterion. A location must pass all criteria to be suitable, and the final rating is suitable or unsuitable.

In our example, we will apply the cartographic model described by the flowchart in Figure 13-4 to a small watershed in a mountainous study area. Application of the refined criteria requires three base data layers — elevation, soils, and roads. For this example we assume the three data layers are available at the required positional and attribute accuracy, clipped to the study area of interest. The need for new data layers often becomes apparent during the process of translating the initial, general criteria to specific, refined criteria, or during the development of the flowchart describing the cartographic model. Once data availability and quality have been assured, we can complete the final flowchart.

Figure 13-10 contains a flowchart of a cartographic model that may identify suitable sites. Spatial data layers are shown as rectangles, and a descriptive data layer name is included within the rectangle. Spatial operations or functions are contained in ellipses, and arrows define the sequence of

data layers and spatial operations. The three base data layers (elevation, soils, and roads), are shown at the top of the flowchart.

There are three main branches in the flowchart in Figure 13-10. The leftmost branch addresses the terrain-related criteria, the center branch addresses the soils criteria, and the right branch applies the road distance criteria. All three branches join in the cartographic model, producing a final suitability classification.

The left branch of the cartographic model is shown in detail in Figure 13-11. This and subsequent detailed figures show a thumbnail of the spatial data layers at each step in the process. Data layer names are adjacent to the spatial data layer. The first two criteria involve terrain-related constraints. Suitable sites are required to possess a restricted set of slopes and aspects. These criteria require slope and aspect data layers, to be calculated and then classified into areas that do and do not meet the respective criteria. The elevation data layer is shown at the top of Figure 13-11; low elevations in black through higher elevations in lighter shades. There are two main river systems in the study area, one running from west to east in the northern portion of the study area, and one running from south to north. Highland areas are found along the north, west, and east margins of the study area.

Slope and aspect are derived from the elevation data layer (Figure 13-11). Lower slope values are shown in light shades, higher slope values are shown in dark shades, and aspects are shown in a range of light to dark shades from 0 to 360 degrees. Slope and aspect layers are reclassified based on the threshold values specified in the criteria listed in Table 13-1. A reclassification table is used to assign values to the slope_suit variable based on the slope layer. Cells with a slope_val less than 30 are assigned a slope_suit of 1, while cells with a slope_val of 30 or higher are given a value slope_suit of 0. Aspect values are also reclassified using a table.

Slope and aspect layers are combined in an overlay, converted from raster to vector, and reclassified to produce a suitable terrain layer (Figure 13-12). Raster to vector conversion is chosen because two of the three base data layers are in a vector format, and because future complex selections might be better supported by the attribute data structure used for vector data sets. This conversion creates polygons that have the attributes

Table 13-1 Original and refined criteria for cartographic model example.

General Criteria	Refined Criteria
Slopes not too steep	Slopes < 30 degrees
Southern aspect preferred	90 < Aspect < 270
Soils suitable for septic system	Specified list of septic-suitable soil units
Far enough from road to provide privacy, but not isolated	300 meter < distance to road < 2,000 meters

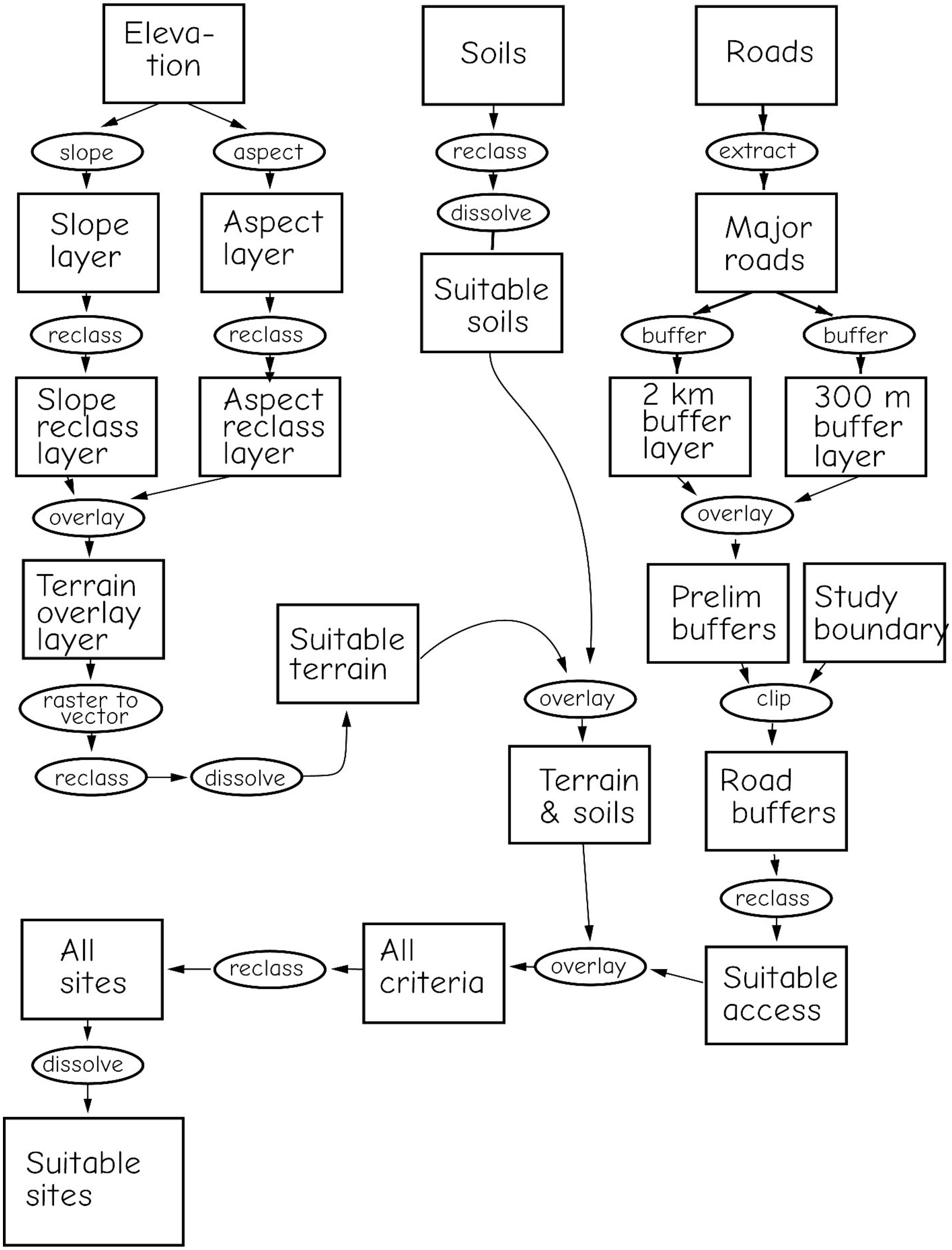
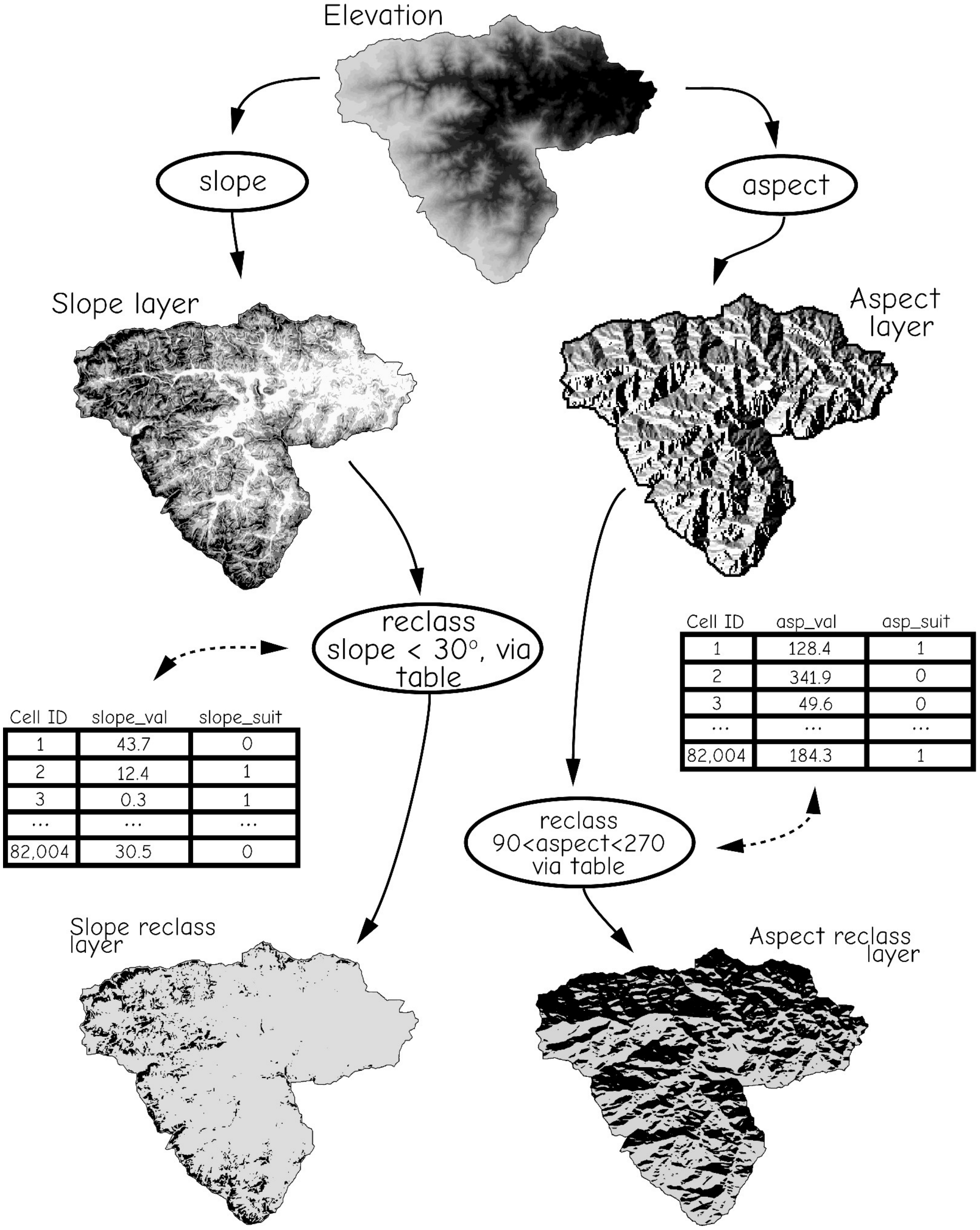


Figure 13-10: Flowchart for the home site suitability cartographic model. Three basic data layers are entered. A sequence of spatial operations is used to apply criteria and produce a map of suitable sites.



Cell ID	slope_val	slope_suit
1	43.7	0
2	12.4	1
3	0.3	1
...	...	...
82,004	30.5	0

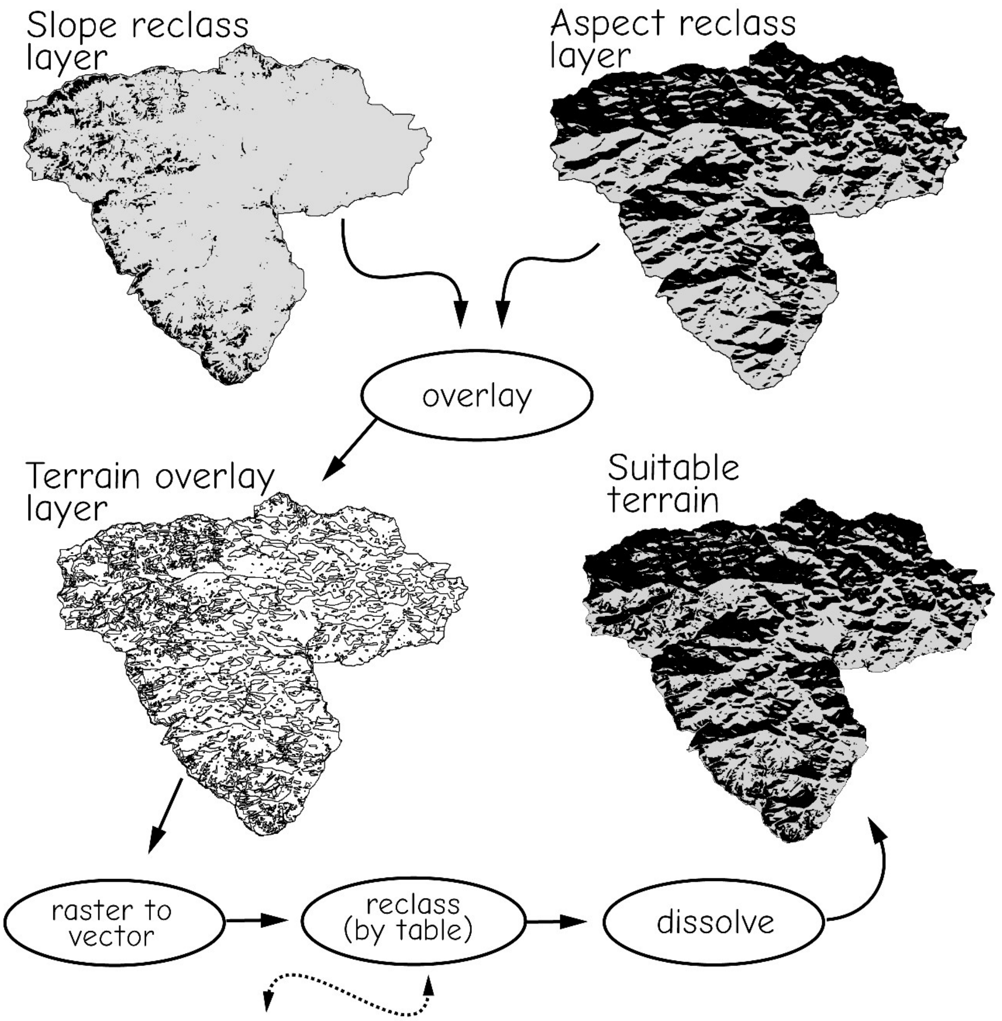
Cell ID	asp_val	asp_suit
1	128.4	1
2	341.9	0
3	49.6	0
...	...	...
82,004	184.3	1

Figure 13-11: A detailed depiction of the leftmost branches of the cartographic model shown in Figure 13-10. Slope and aspect are derived from an elevation data layer for the study region. Both layers are then reclassified using a table assignment. Slope values < 30 are reclassified as suitable (gray), all other slopes as unsuitable (black). Aspect values between 90 and 270 are reclassified as suitable (gray), all others as unsuitable (black).

of the input raster data layer. Note this conversion takes place after the raster layers have been reclassified into a small number of classes, and after the data have been combined to a single layer in an overlay. Raster-to-vector conversion proceeds more quickly after the number of raster classes has been reduced and the data combined in a single terrain-suitability layer.

The terrain overlay must then be reclassified to identify those areas that meet both the slope and the aspect criteria (see the terrain suitability coding in Figure 13-12). Those polygons with a 1 for both slope_suit and asp_suit are assigned a value of 1 for terrain_suit. All others are given a value of 0, indicating they are unsuitable home sites based on the slope and/or aspect criteria.

Because we wish to reduce the number of redundant polygons where possible, a dissolve is applied after the reclassification. This substantially reduces the size of the output data set, and speeds future processing. Reclassified, dissolved terrain data are saved



slope_suit	asp_suit	terrain_suit
0	0	0
0	1	0
1	0	0
1	1	1

The terrain suitability code is assigned based on this reclass table. Only those polygons with suitable slope <u>and</u> aspect are assigned a terrain_suit value of 1.

Figure 13-12: The recoded slope and aspect data layers are combined in an overlay operation, and the result reclassified. Suitable terrain is shown in gray, unsuitable in black.

in a layer labeled Suitable Terrain (Figure 13-12, bottom-right).

The central branch of the cartographic model in Figure 13-10 is shown in Figure 13-13. Digital soil surveys are available that depict homogeneous soil units as polygons. Attribute data are attached to each polygon, including soil type and soil suitability for septic systems. Soils data for the study area may be reclassified based on these septic suitability attributes. A reclassification table assigns a value of 1 to the variable soil_suit if the soil type is suitable for septic systems, 0 if the soil type is not (Figure 13-13).

After reclassification, there may be many adjacent soil polygons with the same soil_suit value. These are grouped using a dissolve operation (data between reclass and

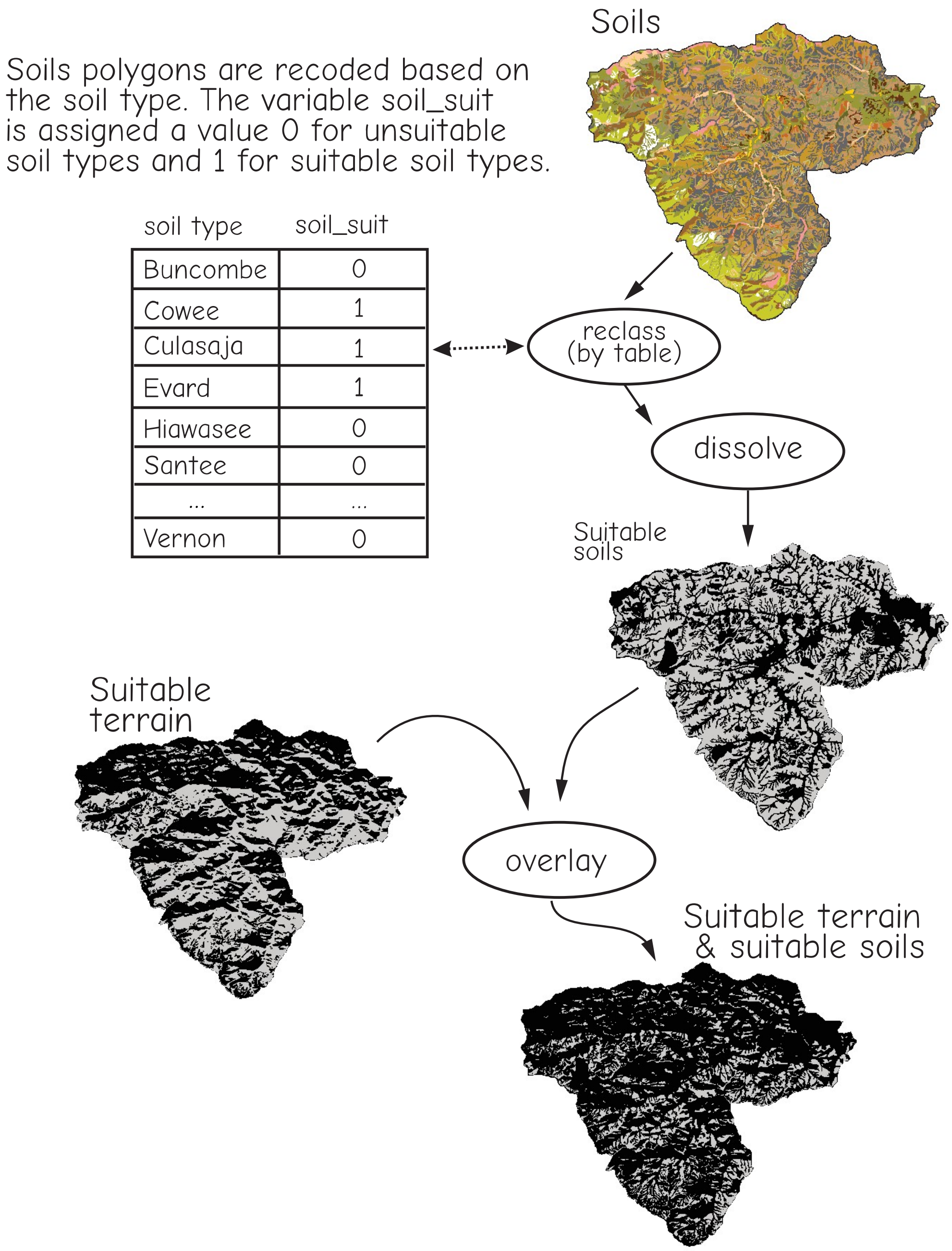


Figure 13-13: A detailed prediction of the center branch of the cartographic model shown in Figure 13-10. Soils data are reclassified into those suitable for septic systems and those not, and then combined with the suitable terrain data layer to identify sites acceptable based on both criteria.

dissolve are not shown in the figure; see Chapter 9 for an example). The dissolve removes boundaries between like polygons, thereby substantially reducing the number of polygons and hence the number of entries in the attribute table. This may be particularly important with complex data sets such as soils data, or with converted raster data, as these often have thousands of entries, many of which will be combined after the dissolve.

The right branch of the cartographic model in Figure 13-10 is presented in Figure 13-14. The Roads data layer is obtained and Major roads extracted. This has the effect of removing all minor roads from consideration in further analyses. What constitutes a major road has been defined prior to this step. In this case, all divided and multi-lane roads in the study area were selected. Two buffers are applied, one at a 300 m distance and one at a

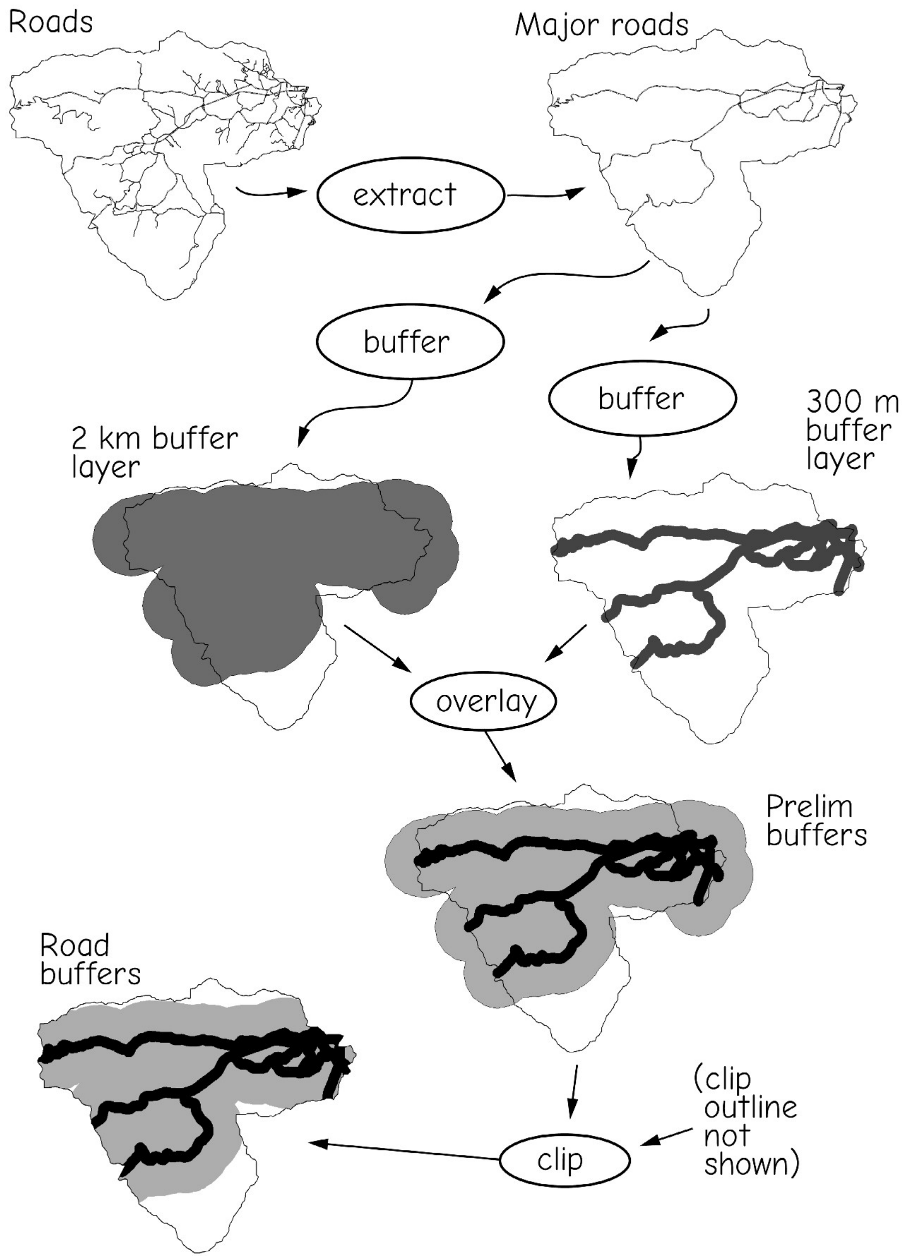
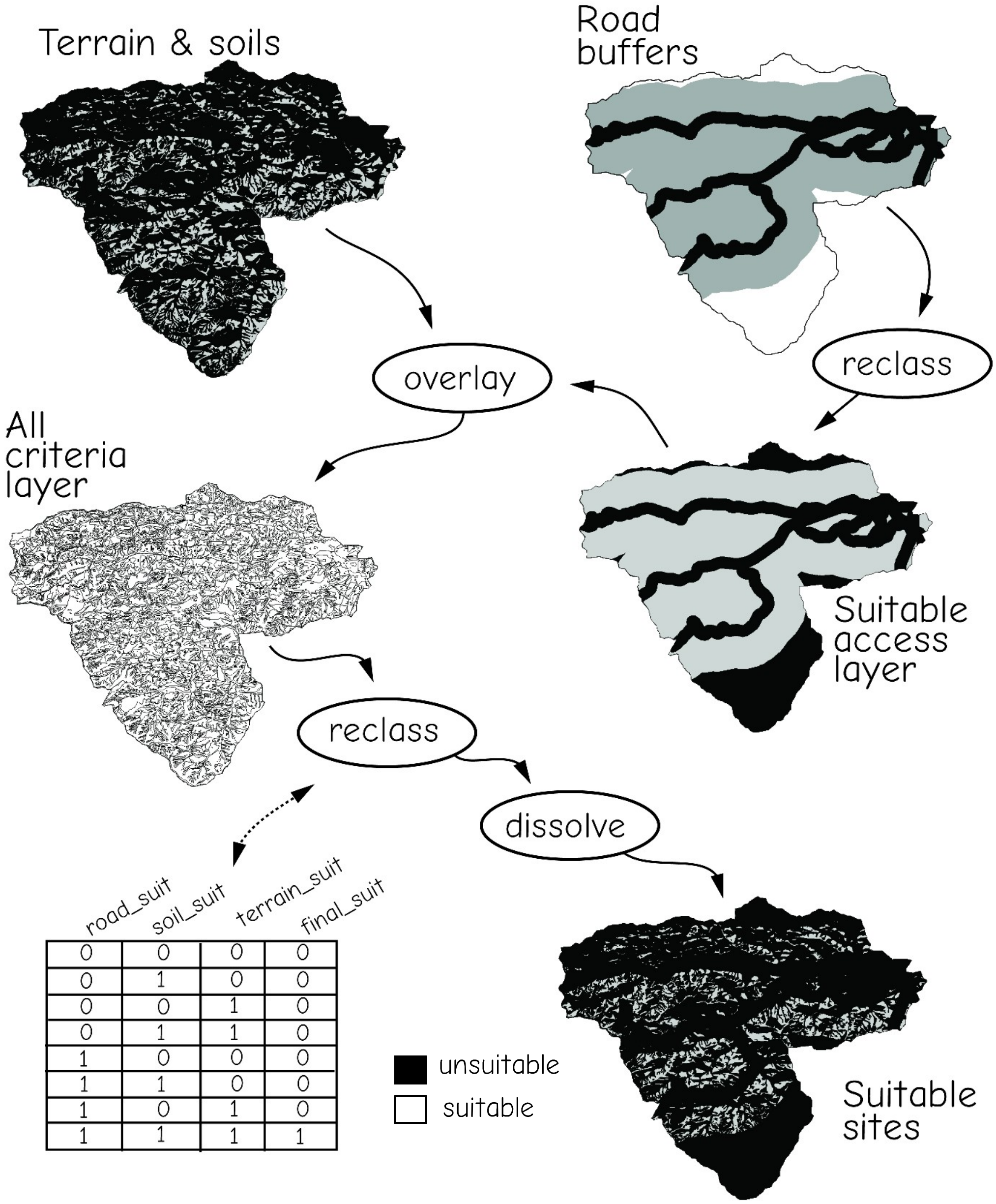


Figure 13-14: A detailed representation of the right branch of the cartographic model shown in Figure 13-10. Roads are buffered at 300 meters and 2 kilometers, and these buffers overlain. The buffers are clipped to the study region, and suitable areas more than 300 meters and less than 2 kilometers from roads identified.

2 km distance from major roads. These buffers are then overlaid. Because the buffer regions extend outside the study area, the buffers must be clipped to the boundary of the study area. These data are then reclassified into suitable and unsuitable areas, resulting in the Road buffers layer (lower left, Figure 13-14).

All data layers are combined in a final set of overlays and reclassifications (Figure 13-15). The Suitable access layer, derived from the roads data and criteria, is combined with the Terrain & soils layer. The All criteria layer contains the required spatial data to identify suitable vs. unsuitable sites. This overlay layer must be reclassified based on the road, soil, and terrain suitability variables, classifying all potential sites into a suitable or unsuitable class. A final dissolve yields the final digital data layer, Suitable sites.



road_suit	soil_suit	terrain_suit	final_suit
0	0	0	0
0	1	0	0
0	0	1	0
0	1	1	0
1	0	0	0
1	1	0	0
1	0	1	0
1	1	1	1

Figure 13-15: The overlay and reclassification of the combined data layers. Terrain, soils, and road buffer data are combined in an overlay. These data are reclassed based on the suitability criteria. A final dissolve is applied to reduce the number of polygons, resulting in a final layer of suitable sites.

This example analysis, while simple and limited in scope, illustrates both the flexibility and complexity of spatial data analysis using cartographic models. The cartographic model was simple because only three input spatial data layers were required, and a small set of spatial data operations were used. Most real analyses use many more data inputs. Reclass, overlay, and other operations were used repeatedly. The modeling is flexible because spatial operations may be tailored to the problem. Finally, this example illustrates the complexity that can be included in cartographic models, as over 20 different instances of a spatial operation were applied, in a defined sequence, resulting in at least 15 intermediate data layers as well as the final result layer.

Scripting and Models

Many softwares provide scripting or programing environments to specify a sequence of spatial operations (Figure 13-16). Examples include ArcGIS Model Builder, the QGIS Batch Modeler and Graphical Modeler, and the GRASS wxGUI. These modeling tools can create a chain of operations that may be saved, re-run, shared, or applied with different input data or parameters. The models may be viewed as a recipe for a set of spatial tools, applied to data, with output from operations used as input in subsequent operations.

The programing environments often allow complex flow, including looping through various iterations of data or parameters, and branching based or termination based on conditions. Scripts are often quite helpful in both saving a cartographic model so that it may be repeated with new data, and for documenting the steps applied in an analysis. Scripts are particularly helpful when processing must be applied repetitively to different instances of the same type of data, e.g., to re-project then re-code hundreds of raster data tiles. This might be rather tedious to complete with a standard graphical user interface, but may take only minutes of user time when incorporated into a script.

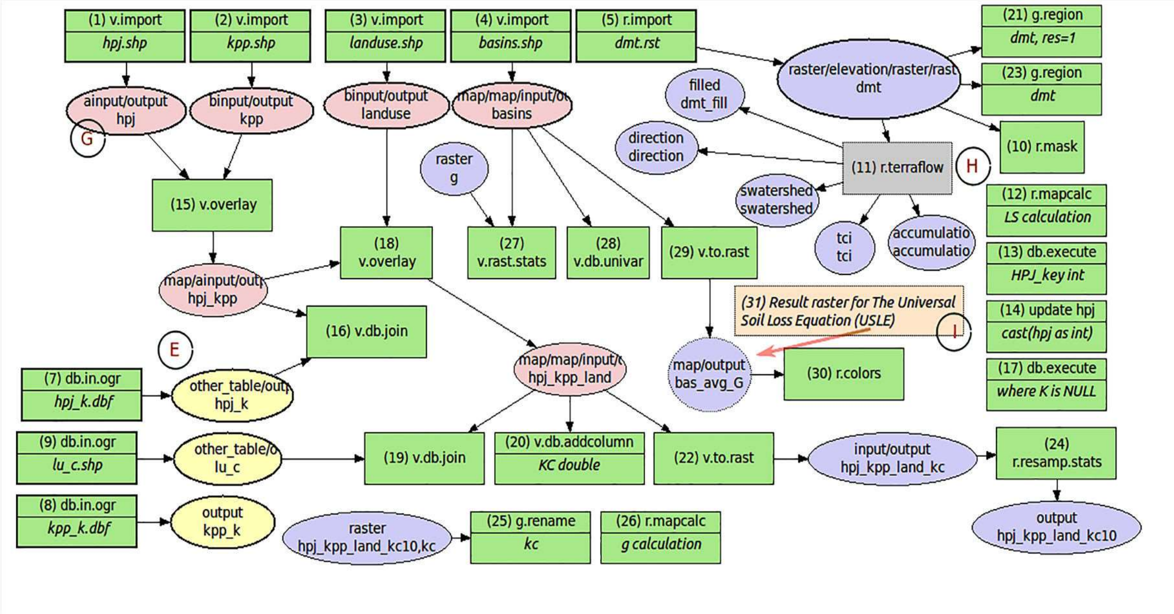


Figure 13-16: An example graphic model developed with the wxGUI for GRASS. Rectangles represent modules or operations ellipses represent data, and arrows show work flow (courtesy OSGEO).

Simple Spatial Models

Predictive spatial models are commonly applied, particularly when there is a well-established model based on point or small-scale observations and analysis, and when the output is a continuous variable, for example, temperature, housing value, soil erosion rates, or cancer frequency.

As noted earlier, our simple spatial models typically are based on one or a few equations, described as:

$$O = f(A, B, C, D, \beta_1, \beta_2, ...) \qquad (13.2)$$

where O is the spatially-reference output; $f()$ is a mathematical operation; A, B, C, D, are variables; and β_i's are equation parameters. For example, NASA has sponsored the development of global models of gross primary productivity (GPP), the total biomass produced globally by plants in any given year. One common model takes the form:

$$GPP = \varepsilon \cdot NDVI \cdot PAR \qquad (13.3)$$

where NDVI is a satellite-based measure of plant abundance, PAR is the amount of sunlight available for photosynthesis, and ε is a conversion efficiency, which may be fixed, or which may depend on additional factors, such as vegetation type or soil dryness or type. In this example, our equation is simple multiplication of the components, and ε is the unique parameter in the simplest case of a fixed ε. In more complicated forms, there is a different ε for each vegetation type, applied accordingly.

Simple spatial models require spatial fields of all variables, and appropriate parameters for all conditions in the modeled area. In our GPP example above, we must have estimates of NDVI and PAR over our prediction region. In this specific case, robust measurements of NDVI have been developed based on repeat satellite measurements, as have methods to estimate PAR from the available meteorology networks and measurement systems. Values of ε have been estimated for dominant vegetation types, and how these parameters vary with other environmental factors like temperature and available moisture. Model estimates have been compared to measurements across a broad range of conditions.

While we call these simple spatial models, as the previous and subsequent examples will show, it is often time consuming and difficult to develop the spatial data and estimate the parameters to apply these models across space. The models are often based on observed relationships and measurements at points or small plots, for example, crop growth on sunny vs. cloudy days, or the change in GPP across nearby forest stands with different NDVI values. These may suffice to estimate ε for the specific types, but differences among vegetation types may require repeat measurements over a broad range of conditions. A network of field stations, perhaps in combination with remotely sensed data, may be required to estimate the input variables, for example, PAR at the required intervals across the landscape.

The Revised Universal Soil Loss Equation (RUSLE) and its precursor the, Universal Soil Loss Equation (USLE), are among the most widely used simple spatial models:

$$E = R \cdot K \cdot C \cdot P \cdot L \cdot S \qquad (13.4)$$

where E is average annual erosion, R is a rainfall factor, K reflects soil erodibility, C integrates crop effects, P accounts for management practices, L reflects slope length, and S represents steepness.

The USLE/RUSLE predict soil erosion on farm fields, and have been under development since the 1930s. Rainfall intensity, soil properties, crop type, slope steepness, and slope length factors have been measured in tens of thousands of plots. Supporting information has been developed for the entire country by the U.S. Natural Resource Conservation Service, including soil and climate factors for the United States, and the impacts of common crop types and manage-

ment regimes. The USLE and RUSLE have been widely applied in other countries.

The RUSLE has been widely applied within a GIS framework for erosion estimates on a catchment or larger scales. The model is relatively simple, much of the input data have been developed and are publicly available, and the outputs are of broad interest. Methods for applying the model have varied, in part because the model was developed for individual fields, but spatial data are often not available on a per-field basis. While the rainfall factor, R, is generally similar across county-sized areas spanning tens of kilometers, other factors often change on a field or subfield basis. Applications often differ in the methods for estimating the management and crop factors, and in particular a combination of slope length and steepness factors.

Estimating driving variables across space often presents choices, as illustrated in the calculation of the RUSLE slope factors, L and S (Figure 13-17). Simple spatial models are often based on small area studies for which all variables may be easily measured. This is often not true when applying the models to larger areas. Slope steepness (S factor) is easily estimated within a raster framework, but slope length (L) is considered uniform at a fixed length of 22.1 m (72.6 ft) in the standard RUSLE. Application of the RUSLE to convergent or divergent slopes or to lengths or cell sizes different than the standard may result in prediction errors. This challenge has been the focus of

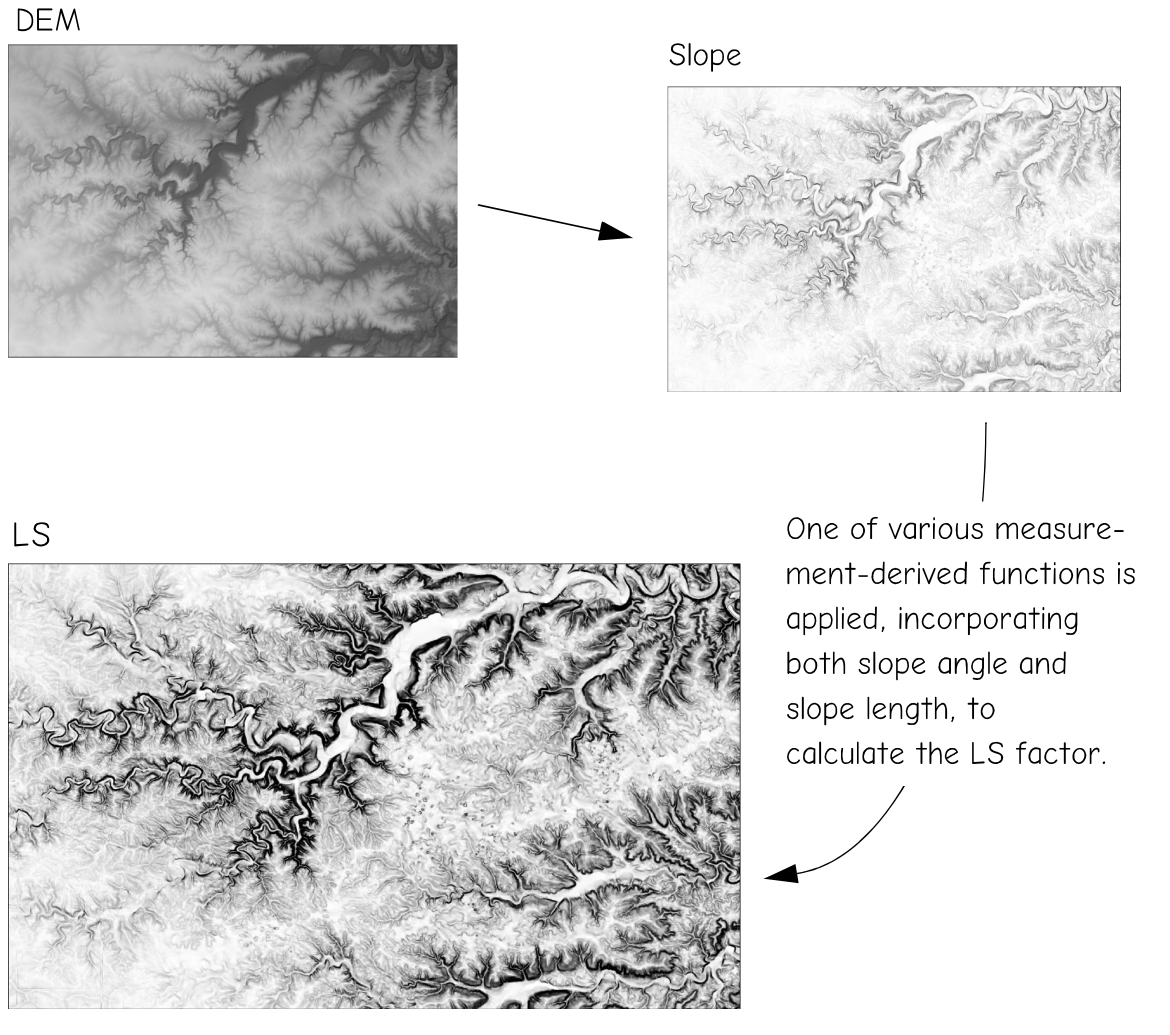


Figure 13-17: Scaling from plots to landscapes requires estimating all input data across space. For the USLE/RUSLE, several methods have been developed for estimating the LS factor, generally based on a combination of slope and slope length derived from DEM data.

many studies, and the book chapter by Wilson and Lorang, listed in the references, describes some of the methods used to effectively estimate a combination of L and S.

Remaining K, C, and P factors may be derived from standard spatial data sets, for example, NASS or NLCD data for landcover/crop type and treatment, and K factors from SSURGO data (Figure 13-18). Application of the model to the spatial data, here in a cell-by-cell multiplication, yields estimates of erosion across a region.

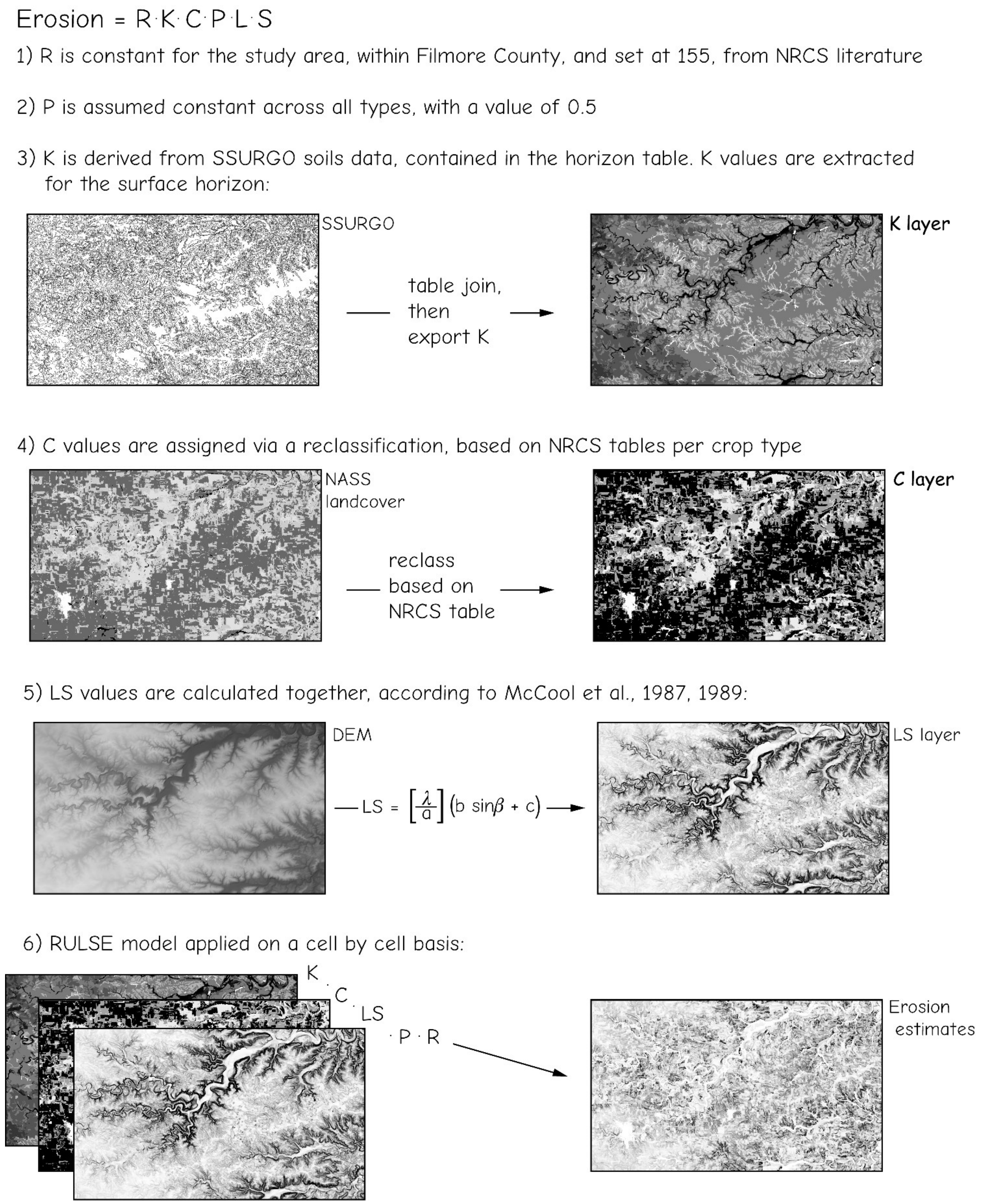


Figure 13-18: USLE/RUSLE erosion estimates may be calculated from appropriately developed base layers.

Spatio-Temporal Models

Spatio-temporal models have been developed and applied in a number of disciplines. This is an active area of both research and application, as there are many fields of study and management that require analysis and predictions of spatially and time varying phenomena. We will briefly discuss some basic characteristics of spatio-temporal models. We will then describe their differences from other models, discuss some basic analysis approaches, and describe two examples of spatio-temporal models.

Spatio-temporal models use spatially explicit inputs to calculate or predict spatially explicit outputs (Figure 13-19). Rules, functions, or some other processes are applied using spatial and often nonspatial data. Input variables such as elevation, vegetation type, human population density, or rainfall may be used as inputs to one or more mathematical equations. These equations are then used to calculate a value for one or more spatial locations. The values are often saved in a spatial data format, such as a layer in a GIS.

Spatio-temporal models involve at least a three-dimensional representation of one or more key attributes – variation in planar (X-Y) space and through time. A fourth dimension may be added if the vertical (Z) direction is also modeled. We arbitrarily treat spatially variable network analyses separately, because networks are constrained to a subset of two-dimensional space. Spatio-temporal models may also be classified by a number of other criteria: whether they treat continuous fields or discontinuous objects, if they are process based or rely on purely fit models, and if they are stochastic or deterministic. Combinations of these model char-

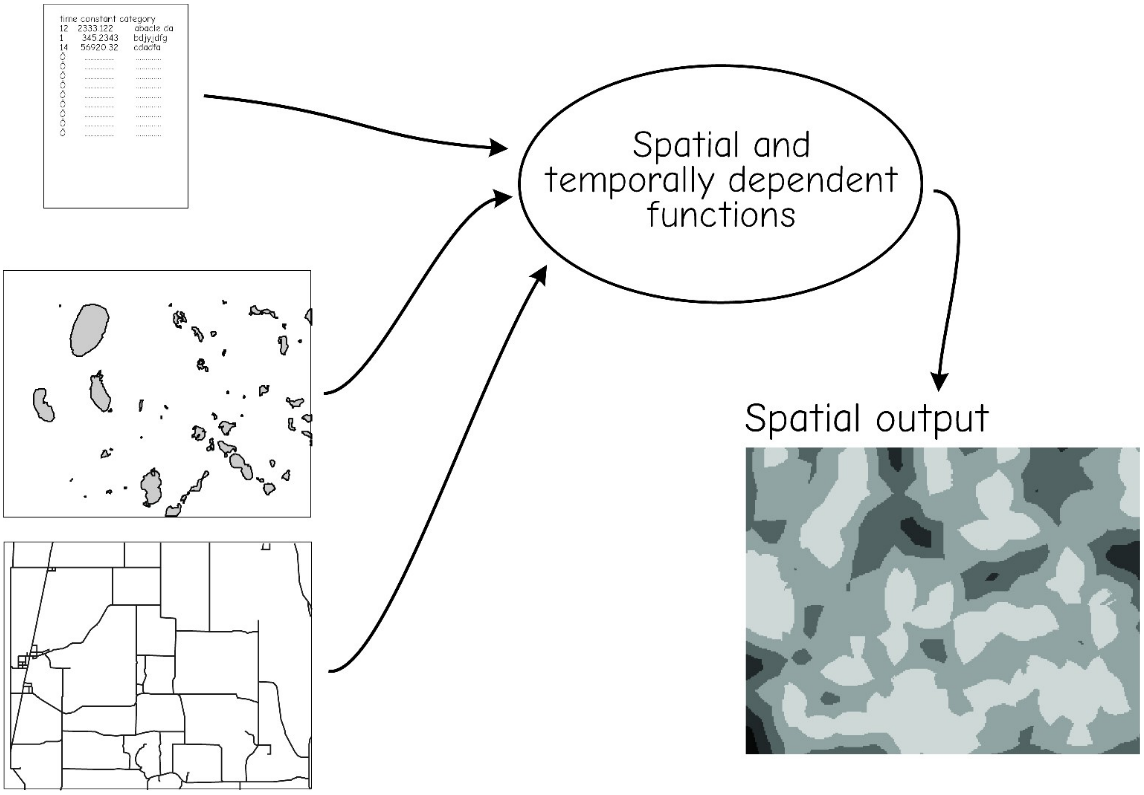


Figure 13-19: Spatio-temporal models combine spatial and aspatial data with time-variant functions to produce spatial output.

acteristics lead to a broad array of spatio-temporal model types.

Models of continuous phenomena predict values that vary smoothly across time or space. Air temperature, precipitation, soil moisture, and atmospheric pollutants are examples of continuous variables that are predicted using spatio-temporal models. Soil moisture this month may depend on soil moisture last month and the temperature, precipitation, and sunshine duration in the intervening period. All these factors may be entered in spatial data layers, and the soil moisture predicted for a set of points.

Models of discrete phenomena predict spatial or attribute characteristics for discontinuous features. Boundaries for vegetation types are an example of features that are often considered discrete. We use a line to identify the separation between two types, for example, between a grassland and a forest. A spatial model may consider the current position of the forest and grassland as well as soil type, fire prevention, and climatic data to predict the encroachment of forest on grassland sites. The boundaries between new forests and grasslands are always discrete, although their positions shift through time.

Models are considered process based if their workings in some way represent a theoretical or mechanistic understanding of the processes underlying the observed changes, and models are purely fit models when they do not. We may predict the amount of water flowing in a stream by a detailed spatial representation of the hydrologic cycle. Many processes may be explicitly represented by equations or subroutines in a spatial model. For example, rainfall location and intensity may be modeled through time for each raster cell in a study area. We can then follow the rainwater as it infiltrates into the soils and joins the stream system through overland flow, subsurface flow, and routing through stream channels. Calculations for these processes may be based on slope, topography, and channel characteristics. These processes are tied together in space. Calculations are performed at each point on the landscape; these calculations increase or decrease water flow or other conditions at adjacent, downslope locations.

Rainfall might be modeled differently using a purely fit, statistical approach. A purely fit model might simply measure precipitation in the previous hour and average the precipitation for the previous week and previous month, and predict stream flow at a point. Processes such as evaporation or subsurface flow are not explicitly represented, and the output may be a statistical function of the inputs. The model may be more accurate than a process-based approach, in that the predicted outflow at any point in the stream may be closer to measured values than those derived from a process-based model. Conversely, the output may be poorer, in that the measurements may be farther from predictions. Process modelers argue that by incorporating the structure and function of the system into a process model we may better predict under new conditions, for example, for extreme drought or rainfall events never experienced before. They also argue that process models aid in our understanding a system and in generating new hypotheses about system function.

Besides being continuous or discrete and process or fit, models may be stochastic or deterministic. A deterministic model provides the same outputs every time it is given exactly the same inputs. If we enter a set of variables into a model without modifying the model, it will always produce exactly the same results. A stochastic model will not. Stochastic models often have random generation or some other variability generation procedures that change model results from run to run, even when using exactly the same inputs.

A disease spread process is a good example of a phenomenon that might be modeled with a stochastic process. Disease may occur at a set of locations, and may be spread through the atmosphere, spread through water, or carried by animals or humans to initiate new disease centers. A doctor might model disease infection and growth stochastically. A random number

might be generated, and the new center started at a location based on this number. The doctor might use another totally or partially random process to control how the new infection center grows or “dies” in the spatial model. Thus, the map of disease locations after different model runs may differ, even though the runs were initiated with identical input conditions.

With most spatial models, the target location of the model output is usually, but not always, the location of the inputs. For example, a demographics model may use a combination of current population in a census tract, housing availability and cost, job opportunities and location, general migration statistics, and age and marital status of those currently in the census tract to predict future population for the census tract. This model has a target location, the census tract, that is the same as the location for most of the input data.

In contrast, the target location of the model outputs may be different than the location of the inputs. Consider a fire behavior model. This model might predict the location of a wildfire based on current fire location, wind speed, topography (fires burn faster upslope than down), and vegetation type and condition. Fire models often incorporate mechanisms to predict fire spread beyond the current burn front of a fire. Embers often are lifted above a fire by the upwelling heated air. These embers may be blown well in advance of a fire front, starting spot fires at some distance away from the main fire. In this case, the target location for a calculation in the spatial model is not the same as the input locations.

Cell-Based Models

Spatial-temporal models often are implemented as *cell-based models*. A cell-based model invokes a set of functions and logic, driven by cell values, to update these or other cell values through time. Input values at a starting time, t_0, may be derived from multiple layers. These input values are entered into functions that calculate the new values for the target layer or layers at the next time step, t_1. The process is then repeated, and the values in the target layer(s) evolve through time (Figure 13-20).

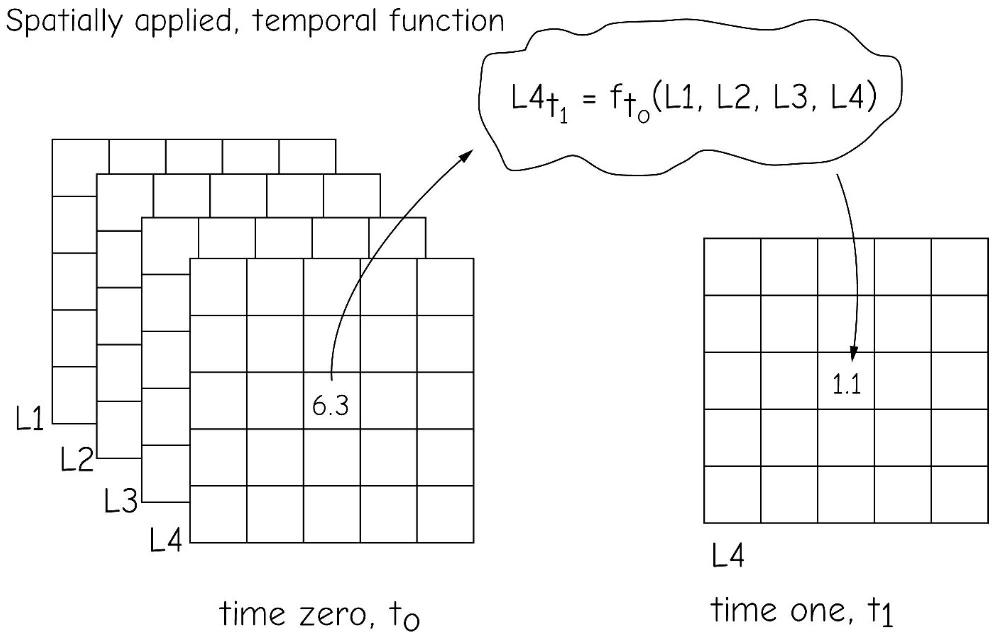


Figure 13-20: Time varying cell-based models use one to multiple input cell values from a starting time, t_0, and update cell values for some layer(s) at time t_1. The process is repeated for a specified length of time or

The previous example of erosion due to surface runoff shows how cell-based modeling can be extended beyond the static analysis of the USLE. Although there are many erosive forces, water is the primary cause of erosion over most of the globe. The amount of soil erosion depends on many factors, including the rainfall rate, how fast the rainwater is absorbed by the soil (permeability), the type of soil, the slope at the site, and how much water is flowing from uphill cells. Some of these properties do not usually change over a rainstorm, for example, slope or soil type, while other features do, such as rainfall rate and downflowing water. All of these factors may be provided as cell-based layers, some that change with time, and some that are static. These layers are then included in an equation to calculate erosion at each cell location for a grid. Rainfall and flow rates may be updated at each step, and the resultant erosion calculated and placed in an output layer, as shown in Figure 13-20.

Example 1: Process-Based Hydrologic Models

Water flows downhill. This simple knowledge was perhaps sufficient until humans began to build houses and roads, and populations grew to dominate most of the Earth's land surface. Land scarcity has led humans to build in low-lying areas, and farming, wetland drainage, and upstream development have all contributed to more frequent and severe flooding.

Water models are needed because demands for water resources are exceeding the natural supply in many parts of the world. Population pressures have driven farms, cities, and other human developments into flood-prone areas; these same developments have increased the speed and amount of rainfall runoff, thereby increasing flood frequency and severity. These factors are spurring the development of spatio-temporal hydrologic models. The models are often used to estimate stream water levels, such that we may better manage water resources and avoid loss of property or life due to flooding.

Many spatio-temporal hydrologic models predict the temporal fluctuations in soil moisture, lake or stream water levels, and discharge in hydrologic networks. The network typically consists of a set of connected rivers and streams, including impoundments such as lakes, ponds, and reservoirs (Figure 13-21). This network typically has a branching pattern. As you move upstream from the main discharge point for the network, streams are smaller and carry less water. Water level or discharge may be important at fixed points in the hydrologic network, at fixed points on land near the network, or at all points in the landscape. The hydrologic network is often embedded in a watershed, defined as the area that contributes downslope flow to the network.

Spatially explicit hydrologic models are almost universally dependent on digital ele-

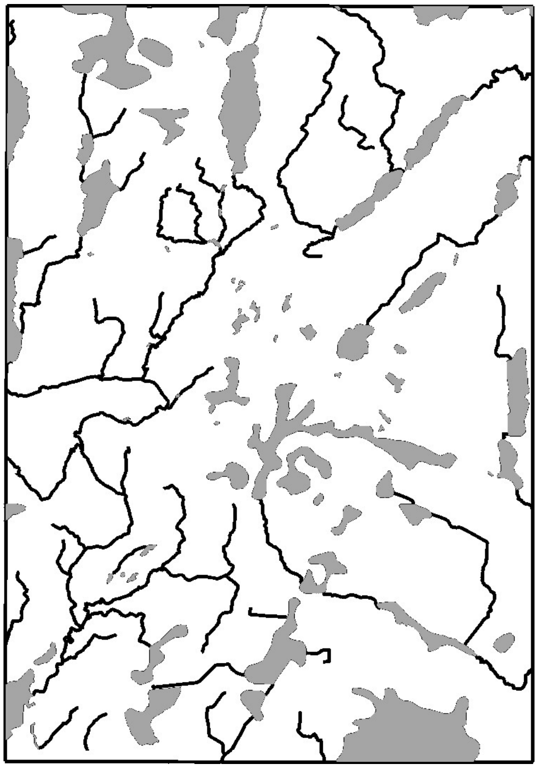


Figure 13-21: An example of a hydrographic network. Lakes and/or rivers form an interconnected network. Water may be routed from upland areas to and through this network.

vation data. DEMs define watershed boundaries, water flow paths, the speed of downslope movement, and stream location (Chapter 12). Slope, aspect, and other factors that effect hydrologic systems may be derived from DEMs. For example, evaporation of surface water and transpiration of soil water depend on the amount of solar radiation. Site solar radiation depends on the slope and aspect at each point, and in mountainous terrain it may also depend on surrounding elevations, due to shading. Site-specific variables representing slope and aspect are used when estimating evaporation or plant use of water.

Slope and aspect are often used to define an important spatial data layer in hydrologic modeling – flow direction. This layer defines the direction of water flow at important points on the surface. If a raster data structure is used, flow direction is calculated for every cell. If a vector data structure is used, flow direction is defined between adjacent or connected vector elements.

Many hydrologic models represent water flow through raster grid cells (Figure 13-22). Water falls on each cell via precipitation. Precipitation either infiltrates into the soil or flows across the surface, depending on the surface permeability at the cell. For example, little water infiltrates for most human-made surfaces, such as parking lots or buildings. These sites have low permeability, so most precipitation becomes surface flow. Conversely, nearly all precipitation infiltrates into most forest soils.

Downslope water flow is also calculated in the model, depending on a number

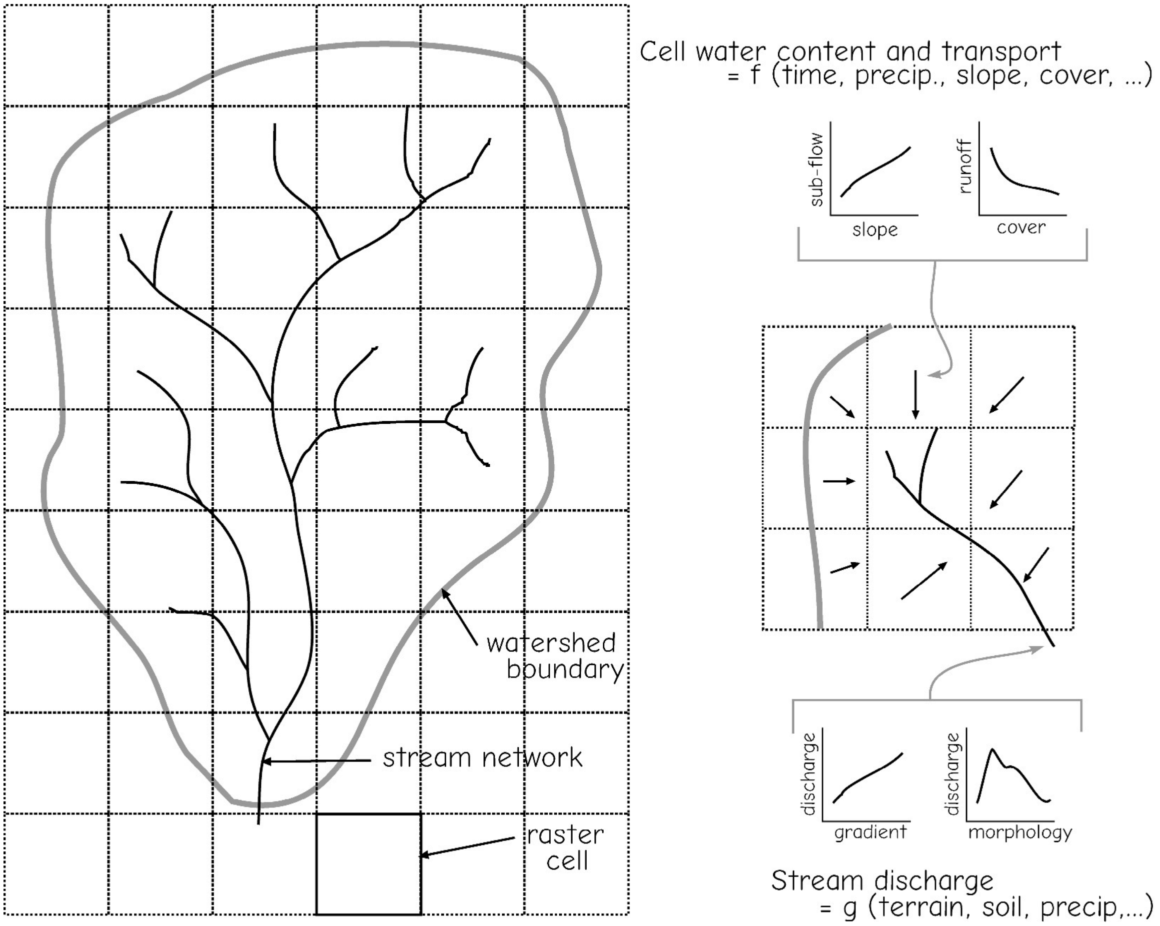


Figure 13-22: Watershed and stream network hydrology may be modeled in a raster environment. Cell characteristics for a watershed are modeled, and water accumulation and flow driven through the system. Soil water, stream levels, and stream discharge depend on spatially and time dependent functions.

of factors at each cell. Slope and flow direction determine the rate at which water flows downhill. Downslope flow eventually reaches the hydrologic network and is routed via the network to the outlet. Mathematical functions describing cell-specific precipitation, flows, and discharge may be combined to predict the flow quantity and water level at points in the watershed and through the network.

Spatio-temporal hydrologic models often require substantial data development. Elevation, surface and subsurface permeability, vegetation, and stream network location must be developed prior to the application of many hydrologic models. DEM data may require substantial extra editing because terrain largely drives water movement. For example, local sinks occur much more frequently in DEMs than in real surfaces. Sinks may occur during data collection or during processing. Sinks are particularly troublesome when they occur at the bottom of a larger accumulation area. Modeled water may flow into the sink but may not flow out, depending on how water accumulation is modeled, while on the real earth surface the water may flow freely downhill. Local spikes in the model may push water incorrectly to surrounding cells, although they typically cause fewer problems than sinks. Both sinks and spikes must be removed prior to application of some hydrologic models.

Example 2: LANDIS, a Stochastic Model of Forest Change

Many human or natural phenomena are analyzed through spatially explicit stochastic models. Disease spread, the development of past societies, animal movement, fire spread, and a host of other important spatial phenomena have been modeled. All these phenomena have a random element that substantially affects their behavior. Events too obscure or complex to predict may cause large changes in the system action or function. For example, wind speed or dryness on a given day dramatically affects fire spread, yet wind speed is notoriously difficult to predict. Spatially explicit, stochastic models allow us to analyze the relative importance of component inputs and processes, and the nature and variability of system response. Is it stochastic variation in wind, fuel amount, or fuel type that is most responsible for the variable nature of fire spread? We will discuss one spatial stochastic model — LANDIS (LANdscape DISturbance) — that incorporates techniques used in a wide range of models (Figure 13-23).

Forest vegetation changes through time. Change may be caused by the natural aging and death of a group of trees, replacement by other species, or due to periodic disturbances such as fire, windstorms, logging, insects, or disease outbreaks. Because trees are long-lived organisms, the composition and structure of forests often change on temporal scales exceeding a normal human life span. Human actions today may substantially alter the trajectory of future change. We often need to analyze how past actions have led to current forest conditions, and how present actions will alter future conditions.

Forest disturbance and change are important spatial phenomena for many reasons. Humans are interested in producing wood and fiber, preserving rare species, protecting clean water supplies and fish spawning areas, protecting lives and property from wildfires, and enjoying forest-based recreation.

Forest change is inherently a spatial phenomenon. Fires, diseases, and other disturbances travel across space. The distribution of current forests largely affects the location and species composition of future forests. Seeds disperse through space, aided by wind and water or carried by organisms. Physical and biotic characteristics that largely determine seed and seedling survival, and subsequent forest growth, are variable in space. Some plants are better adapted to grow under existing forests, while others are aided by disturbances that open the canopy. Some species change soil or understory conditions in ways that prevent other species form growing beneath them. Plant succes-

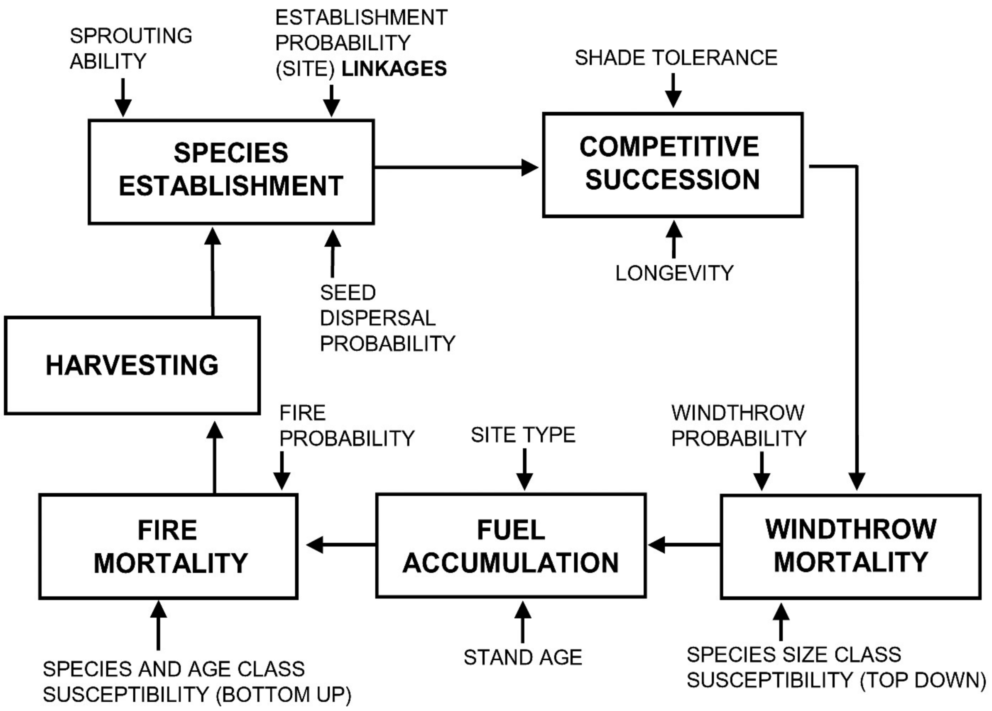


Figure 13-23: The major processes represented in LANDIS, a spatio-temporal forest succession model (courtesy D. Mladenoff).

sion, the replacement of one group of plants or species by another through time, is substantially affected by the current forest distribution and structure. It is not surprising that many process-based models of forest change incorporate spatial data.

Forests are extremely heterogeneous in space, and this complicates our understanding and predictions of forest change. Tree species, size, age, soils, water availability, and other factors change substantially over very short distances. Each forest stand is different, and we struggle to represent these differences. Given the long time scales, broad spatial scales, and inherent spatial variability of forests, many organizations have developed models based on spatial data, models that are in some way integrated into GIS.

LANDIS is an example of a spatially explicit, process-based forest dynamics model. LANDIS has been developed by Dr. David Mladenoff and colleagues, and has been applied to forest biomes across the globe. LANDIS incorporates natural and human disturbances with models of seed dispersal, plant establishment, and succession through time to predict forest composition over broad spatial scales and for its long temporal scales. LANDIS is notable for the broad areas it may treat at relatively high resolution, and long temporal scales. LANDIS has been used to model forest dynamics at a 30 m resolution, over tens of thousands of hectares, and across five centuries.

LANDIS integrates information about forest disturbance and succession to predict changes in forest composition (Figure 13-23). Succession is the replacement of species through time. Succession is common in forests, for example, when fast-growing, light-demanding tree species colonize a disturbed

site, and are in turn replaced by more shade-tolerant, slower growing species. These shade-tolerant species may be self-replacing in that their seeds germinate and seedlings survive and grow in the dense shade. Small gaps from canopy damage or tree deaths allow small patches of light to reach these shade-tolerant seedlings, enabling them to reach the upper canopy. This self replacement can result in a stable, same-species stand over long time periods. This cycle may be broken due to fire, windthrow, logging, or other disturbance event that opens up a stand to a broader range of species. LANDIS simulates large, heterogeneous landscapes, incorporates the interactions of dominant tree species, and includes spatially explicit representations of ecological interactions.

LANDIS Design Elements

The design of LANDIS is driven by the overall objectives for the model, simulating forest disturbance and succession through time. LANDIS also satisfies a number of other requirements. LANDIS readily integrates satellite data sets and other appropriate spatial data, and it simulates the basic processes of disturbance, stand development, seed dispersal, and succession in a spatially explicit manner.

LANDIS is an object-oriented model. Specific features or processes are encapsulated in objects, and object-internal processes are isolated as much as possible from other portions of the model. As an example, there is a SPECIE object that encapsulates most of the important information and processes for each tree species included in the model. Each instance of a SPECIE has a name, for example, "Aspen," and other characteristics such as longevity, shade tolerance, or age to maturity, as well as methods for birth, death, and other actions or characteristics. Because these characteristics and processes are encapsulated in a SPECIE object, they may be easily changed as new knowledge become available. Many models are incorporating this object-oriented design, because it simplifies maintenance and modifications.

LANDIS uses a raster data model that eases the entry of classified satellite imagery, elevation, and other data sets reflecting short-range environmental and forest species variation. Interactions such as seed dispersal, competitions, and fire spread are explicitly modeled for each species occupying each grid cell.

LANDIS tracks the presence of age classes (cohorts) for a number of species in each cell and through time. The model begins with an initial condition: the distribution of species by age class across the landscape. Ten-year age classes are currently represented. The longevity, age of initial seed production, seed dispersal distance, shade tolerance, fire tolerance, and ability to sprout from damaged stumps or roots is recorded for each species. On undisturbed sites, cohorts pass through time until they reach their longevity. Older cohorts "die" and disappear from the cell. Younger cohorts may then appear, depending on the availability of seed.

The spatially explicit representation of seed sources and dispersal is an improvement of LANDIS over many earlier forest succession models. Previous models typically assumed constant or random seed availability. LANDIS is representative of spatially explicit models, in that the specific locations of a process affect that process. Disturbed sites may be occupied by seedlings from a disturbed cell or nearby cells, or by sprouting from trees in a cell prior to disturbance. Cells cycle through the species establishment, succession, disturbance, and mortality processes (Figure 13-24).

The effects of site characteristics on species establishment and interactions are also represented in LANDIS. For example, establishment coefficients are used to represent the interaction between site characteristics and species establishment. Establishment coefficients vary by land type. Fire severity also varies by land type, as may seedling survival. Elevation, aspect, soils, and other

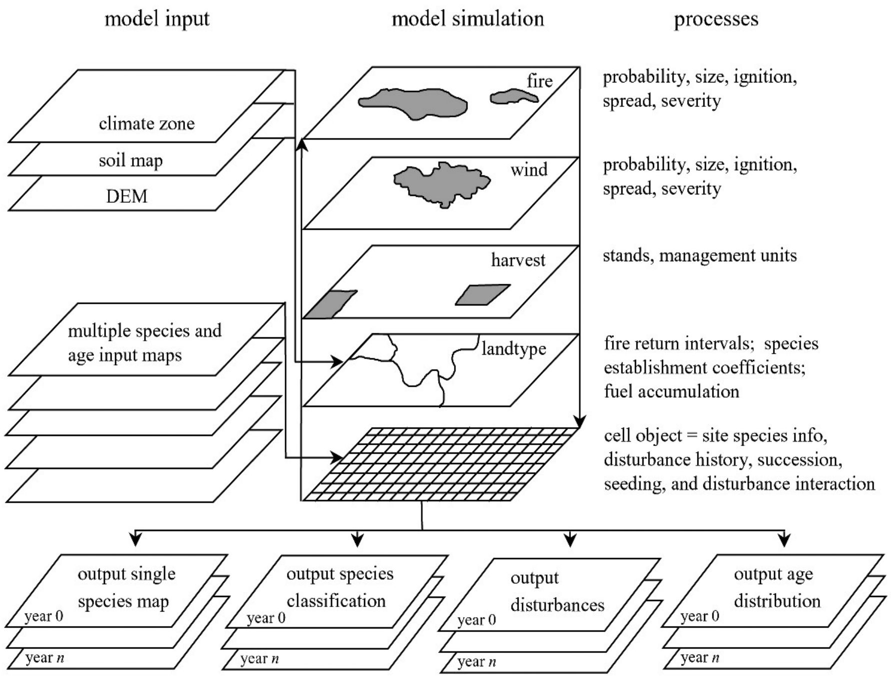


Figure 13-24: Basic spatial data and processes represented in LANDIS.

spatial data are used as input to the spatial model.

Fire and wind disturbances are simulated based on historical records of disturbance sizes, frequencies, and severities. Disturbances vary in these properties across the landscape. For example, wind disturbances may be more frequent and severe on exposed ridges, and fires less frequent, less intense, and smaller in wetlands. Disturbances are stochastically generated, but the variability depends on landscape variables, for example, fires are generated more frequently on dry upland sites.

LANDIS has been applied to a number of forest science and management problems, including the effects of climate change on forest composition and production, the impacts of changing harvesting regimes on landscape patterns, and regional forest assessments (Figure 13-25).

Hundreds of other spatially explicit, temporally dynamic models have been developed, and many more are currently under development. As spatial data collection technologies improve and GIS systems become more powerful, spatio-temporal models are becoming standard tools in geographic science, planning, and in resource management.

Summary

Spatial analysis often involves the development of spatial models. These models can help us understand how phenomena or systems change through space and time, and they may be used to solve problems. In this chapter we described cartographic models, and spatio-temporal models.

Cartographic models often combine several data layers to satisfy a set of criteria. Data layers are combined through the application of a sequence of spatial operations, including overlay, reclassification, and buffering. The cartographic model may be specified with a flowchart, a diagram representing the data layers and sequence of spatial operations. Cartographic models are static in time relative to the other model types.

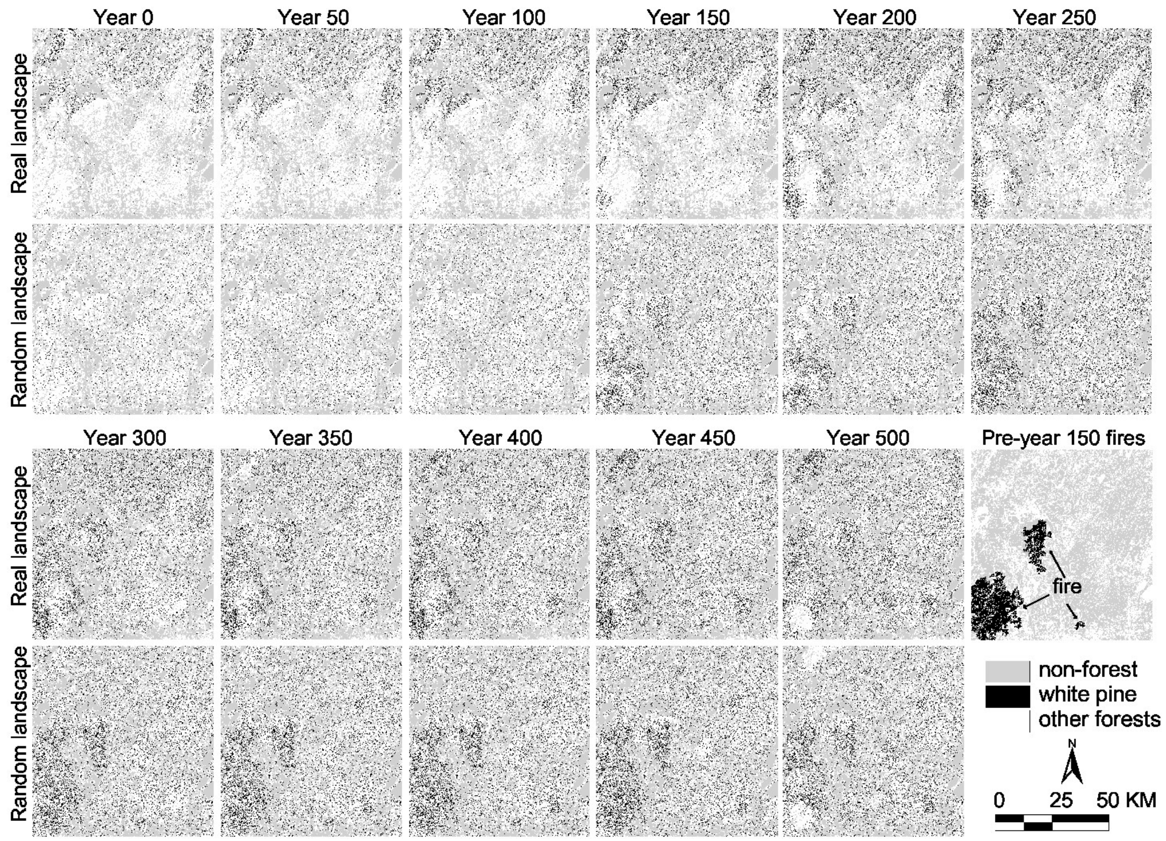


Figure 13-25: Changes in the spatial distribution of white pine, a forest tree species, through time as predicted by LANDIS. This graphic exemplifies the prediction of a feature of interest both spatially and temporally, and is representative of many analytical tools in use or under development.

Spatio-temporal models explicitly represent the changes in important phenomena through time within the model. These models are typically more detailed, and less flexible than cartographic models, in part because spatio-temporal models often include some representation of process. For example, many spatio-temporal models have been developed to model the flow of water through a region, and these models incorporate equations regarding the physics of water transport movement. Models may be stochastic or deterministic, process based or statistical, or they may have a combination of these characteristics.

Suggested Reading

Anselin, L., Syabri, I., Kho, Y. (2006). GeoDA: an introduction to spatial data analysis. *Geographical Analysis*, 38:5–22.

Brady, M., Irwin, E. (2011). Accounting for spatial effects in economic models of land use: recent developments and challenges ahead. *Environmental and Resource Economics*, 48:487-509.

Brown, D., Riolog, R., Robinson, D.G., North, M., Rand, W. (2005). Spatial processes and data models: Towards integration of agent-based models and GIS. *Journal of Geographical Systems*, 7:25–47.

Burrough, P.A., McDonnell, R.A. (1998). *Principles of Geographical Information Systems*. Oxford: Oxford University Press.

Carlson, S. (2000). The amateur scientist: Boids of a feather flock together. *Scientific American*, 283:112–114.

Clarke, K.C., Hoppen, S., Gaydos, L. (1997). A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning*, 24:247–261.

Cliff, A.D., Ord, J.K. (1981). *Spatial Processes: Models and Applications*. London: Pion.

Cooke, W.H.K., Katarzyna, G., Wallis, R.C. (2006). Avian GIS models signal human risk for West Nile virus in Mississippi. *International Journal of Health Geography*, 5:36, doi:10.1186/1476–072X–5–36.

Dixon, B., Uddameri, V. (2016). *GIS and Geocomputation for Water Resource Science and Engineering*. Hoboken: Wiley.

Fotheringham, S., Wegener, M. (2000). *Spatial Models and GIS: New Potential and New Models*. London: Taylor and Francis.

Goodchild, M.F., Steyaert, L.T., Parks, B.O. (1996). *GIS and Environmental Modeling: Progress and Research Issues*. Fort Collins: GIS World Books.

Griffith, D.A., Layne, L.J. (1999). *A Casebook for Spatial Statistical Data Analysis*. Oxford: Oxford University Press.

He, H.S., Mladenoff, D.J., Boeder, J. (1999). An object-oriented forest landscape model and its representation of tree species. *Ecological Modeling*, 119:1–19.

Horn, M.E.T. (2004). Modelling and assessment of demand-responsive passenger transport services. J. Stillwell and G. Clarke (Ed.), *Applied GIS and Spatial Analysis*. Wiley: New York.

Huevelink, G.B.M., Burrough, P.A. (1993). Error propagation in cartographic modelling using Boolean logic and continuous classification. *International Journal of Geographical Information Systems*, 7:231–246.

Jetten, V., Govers, G., Hessel, R. (2003). Erosion models: quality of spatial predictions. *Hydrologic Processes*, 17:887–900.

Johnston, C. (1998). *GIS in Ecology*. Boston: Blackwell Scientific.

Kaufmann, A. (1975). *Introduction to the Theory of Fuzzy Subsets*. New York: Academic Press.

Klir, G.J., Folger, T.A. (1988). *Fuzzy Sets, Uncertainty, and Information*. Englewood Cliffs: Prentice Hall.

Krzanowski, R., Raper, J. (2001). *Spatial Evolutionary Modelling*. Oxford: Oxford University Press.

Malczewski, J.C. (1999). *GIS and Multicriteria Decision Analysis*. New York: Wiley.

McCool, D.K., Brown, L.C., Foster, G.R., Mutchler, C.K., Meyer, L.D. (1987). Revised slope steepness factor for the Universal Soil Loss Equation. *Transactions of the American Society of Agricultural Engineers*, 30:1387–1396.

McCool, D.K., Foster, G.R., Mutchler, C.K., Meyer, L.D. (1989). Revised slope length equation for the Universal Soil Loss Equation, *Transactions of the American Society of Agricultural Engineers*, 32:1571–1576.

Mladenoff, D.J., He, H.S. (1999). Design, behavior and application of LANDIS, an object-oriented model of forest landscape disturbance and succession, In D.J. Mladenoff, D.J., Baker W.L. (Eds.), *Advances in Spatial Modeling of Forest Landscape Change: Approaches and Applications*. Cambridge: Cambridge University Press.

Monmonnier, M. (1993). *How To Lie With Maps*. Chicago: University of Chicago Press.

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A. (1993). Soil attribute prediction using terrain analysis. *Soil Science*, 57:443–452.

Parent, O., LeSage, J.P. (2010). A spatial dynamic panel model with random effects applied to commuting times. *Transportation Research Part B: Methodological,* 44:633–645.

Pinske, J., Slade, M.E. (2010). The future of spatial econometrics. *Journal of Regional Science*, 50:103–117.

Reynolds, C.W. (1987). Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics*, 21:25–34.

Rossiter, D.G. (1996). A theoretical framework for land evaluation. *Geoderma*, 72:165–190.

Running, S.W., Nemani, R.R., Heinsch, F.A., Zhao, M., Reeves, M., Hashimoto, H. (2004). A continuous satellite-derived measure of global terrestrial primary production. *Bioscience*, 54:547–560.

Stillwell, J.A., Clarke, G. (2004). *Applied GIS and Spatial Analysis*. New York: Wiley.

Turner, M.G., Gardener, R.H. (Eds.). (1991). *Quantitative Methods in Landscape Ecology*. New York: Springer Verlag.

Wagner, D.F. (1997). Cellular Automata and Geographic Information Systems. *Environment and Planning*, 24:219–234.

Wilson, J., Gallant, J. (Eds.) (2000). *Terrain Analysis: Principles and Applications.* New York: Wiley.

Wilson, J.P., Lorang, M.S. (2000). Spatial models of soil erosion and GIS, in *Spatial Models and GIS*, Fotheringham, A.S., Wegener, M. (eds.). London: Taylor & Francis.

Wolfram, S. (1994). *Cellular Automata and Complexity*. Reading: Addison-Wesley.

Study Questions

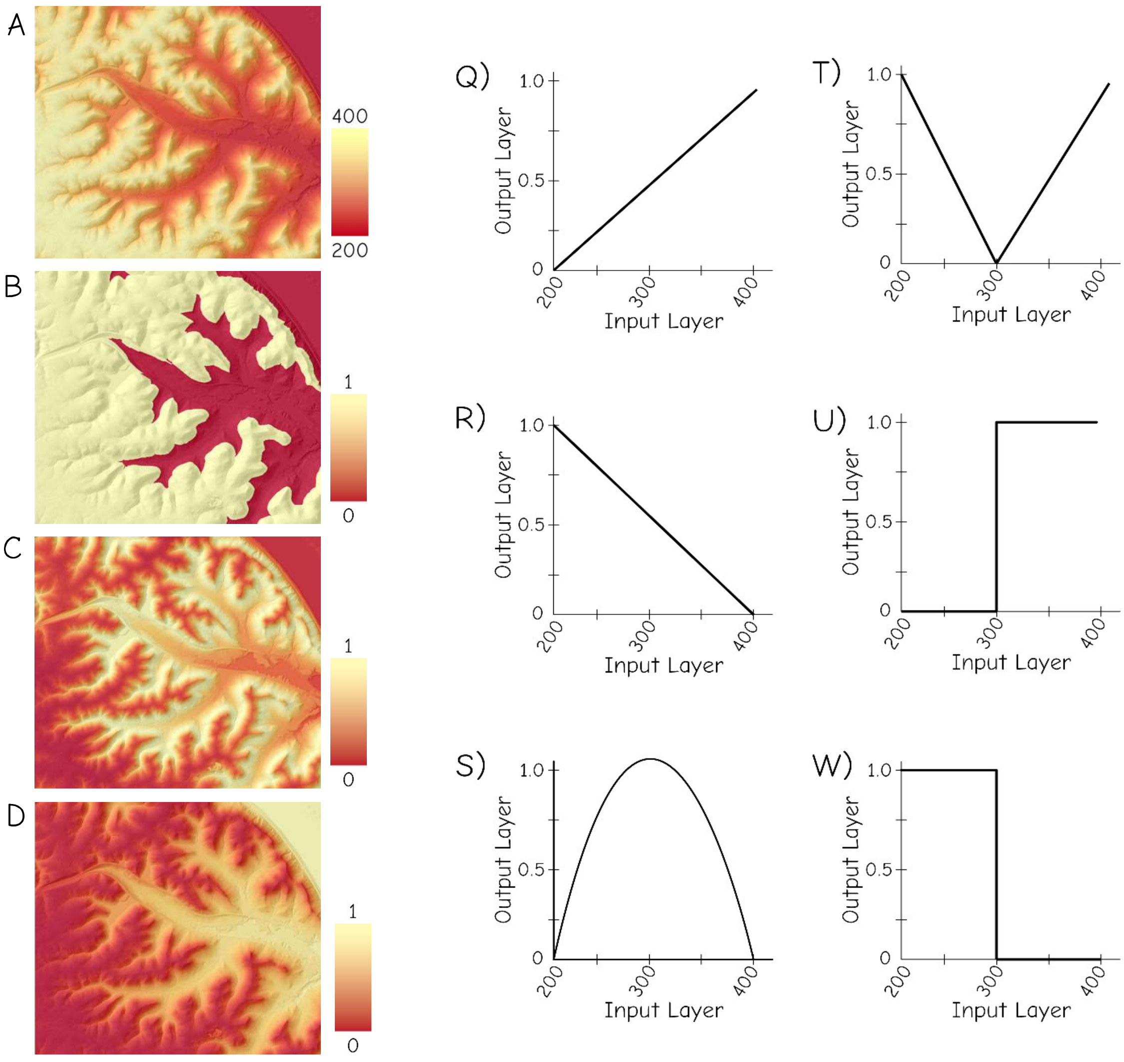
13.1 - Provide an example of a cartographic model, including the criteria and a flow-chart of the steps used to apply the model.

13.2 - Why must the criteria be refined in many cartographic modeling processes?

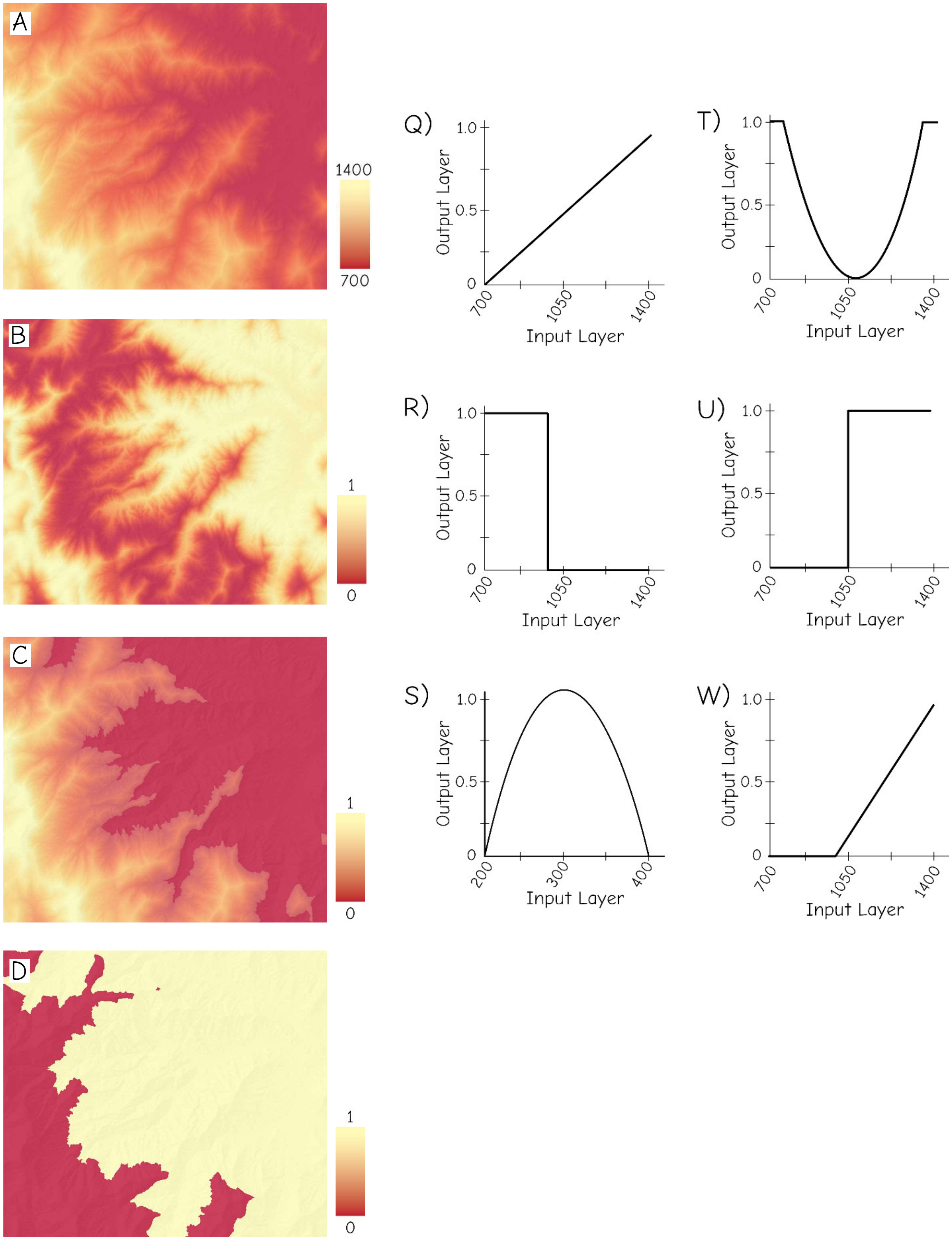
13.3 - What do we mean when we say that most cartographic models are temporally static?

13.4 - What is a discrete vs. continuous weighting in an input layer when combining layers in a cartographic overlay? How do you develop a reasonable continuous ranking function, that is, justify the shape of the curve vs. the level of the input variable?

13.5 - Match the output layers B, C, and D, to the appropriate reclassification graphs Q - W when applied to the original DEM, A. Note that a hillshade surface is superimposed to aid in visualization.



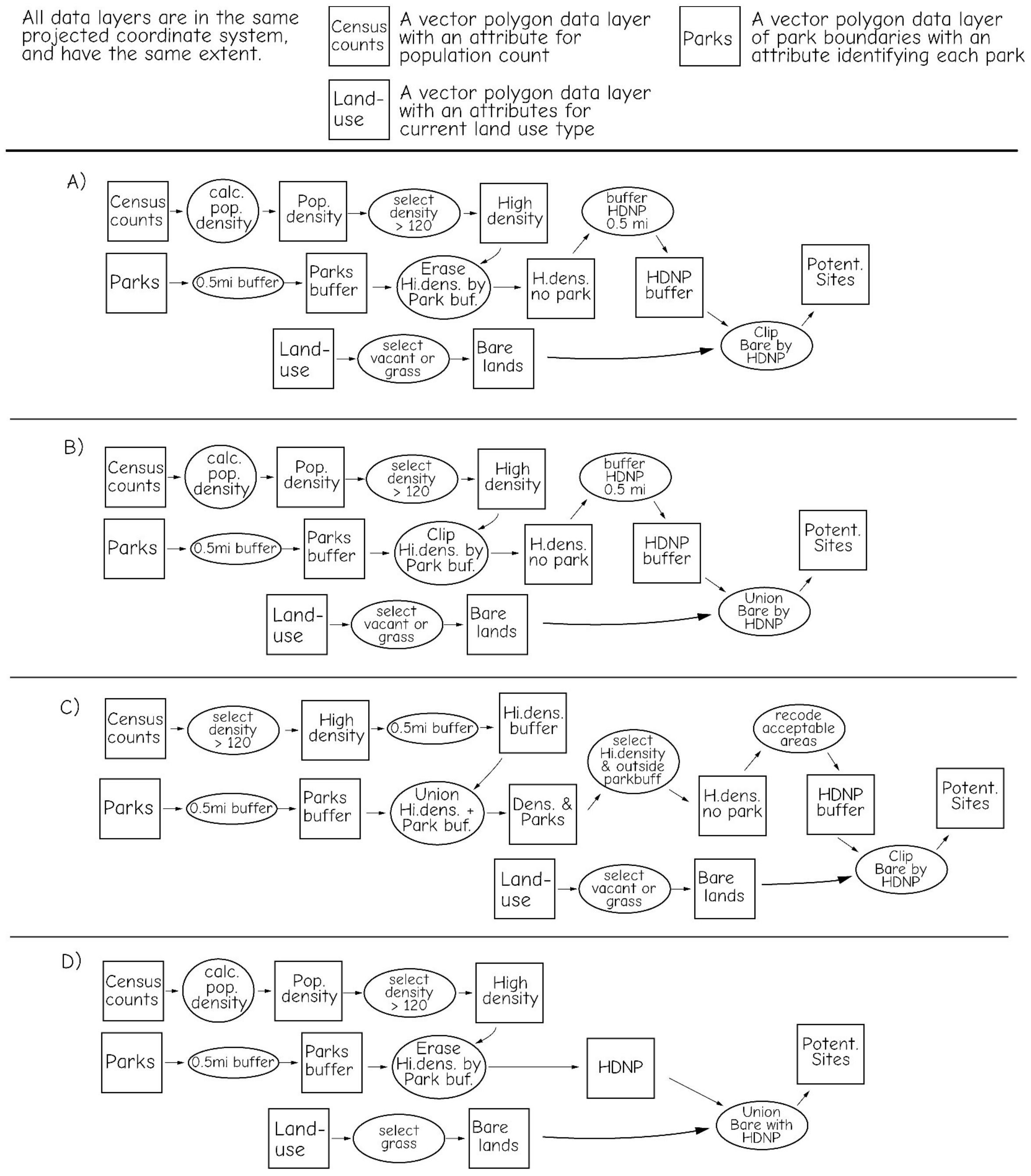
13.6 - Match the output layers B, C, and D, to the appropriate reclassification graphs Q - W when applied to the original DEM, A. Note that a hillshade surface is superimposed to aid in visualization.



13.7 - The figure below depicts four flowcharts of cartographic models to find areas most suitable for a new park. Sites are preferred that meet all of the following criteria:

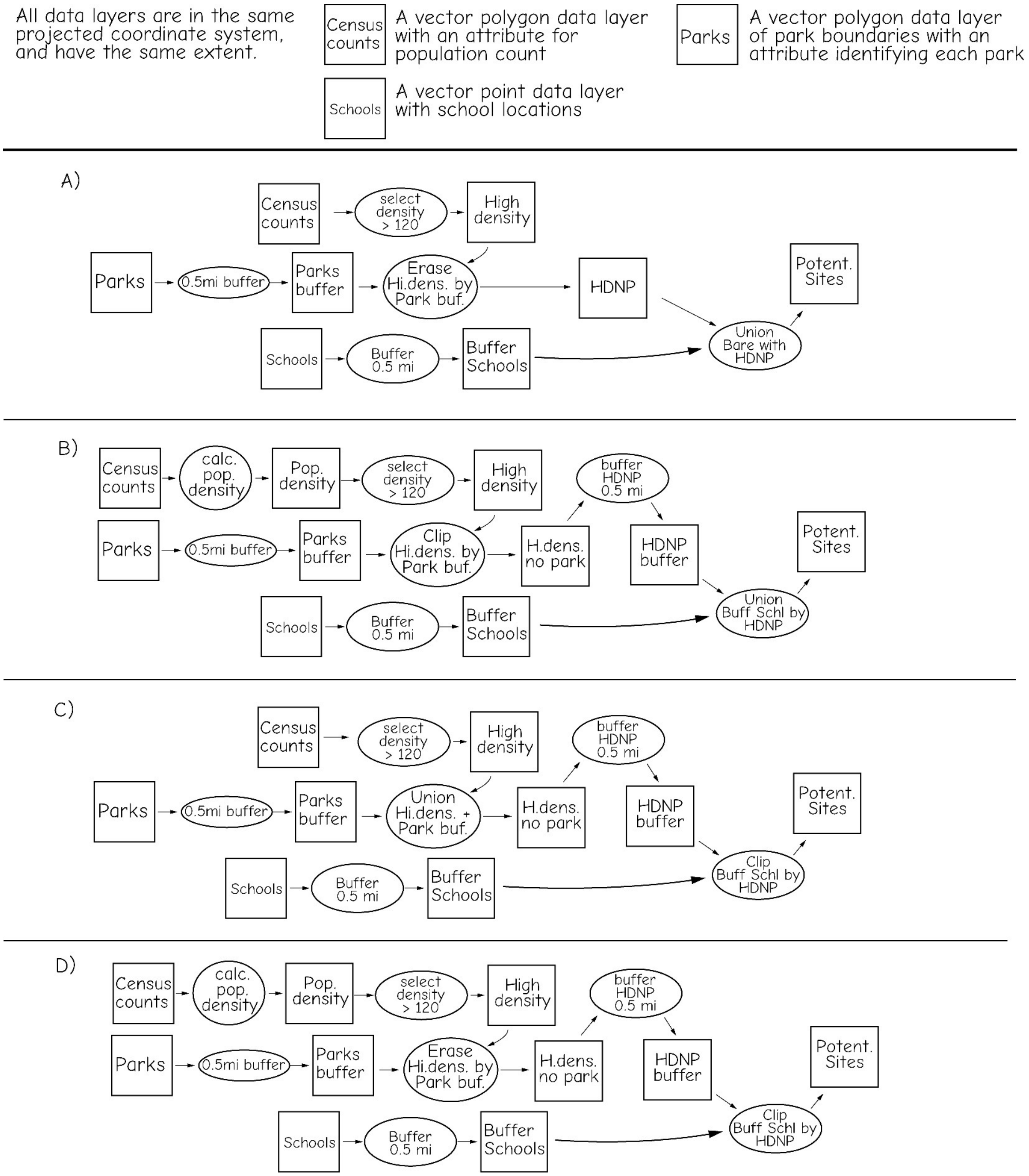
Within 0.5 miles of Census polygons with a density of more than 120 persons per square mile;
Greater than 0.5 miles from an existing park;
Current land use of grass or vacant.

Select the flowchart which best approximates the proper analysis, given the described data. For each other flowchart, list at least one primary way it is inferior to the chosen method. Note that some minor intermediate steps are omitted/subsumed into operations for all flowcharts, so do not cite a step omitted in both the best and alternate flowcharts.



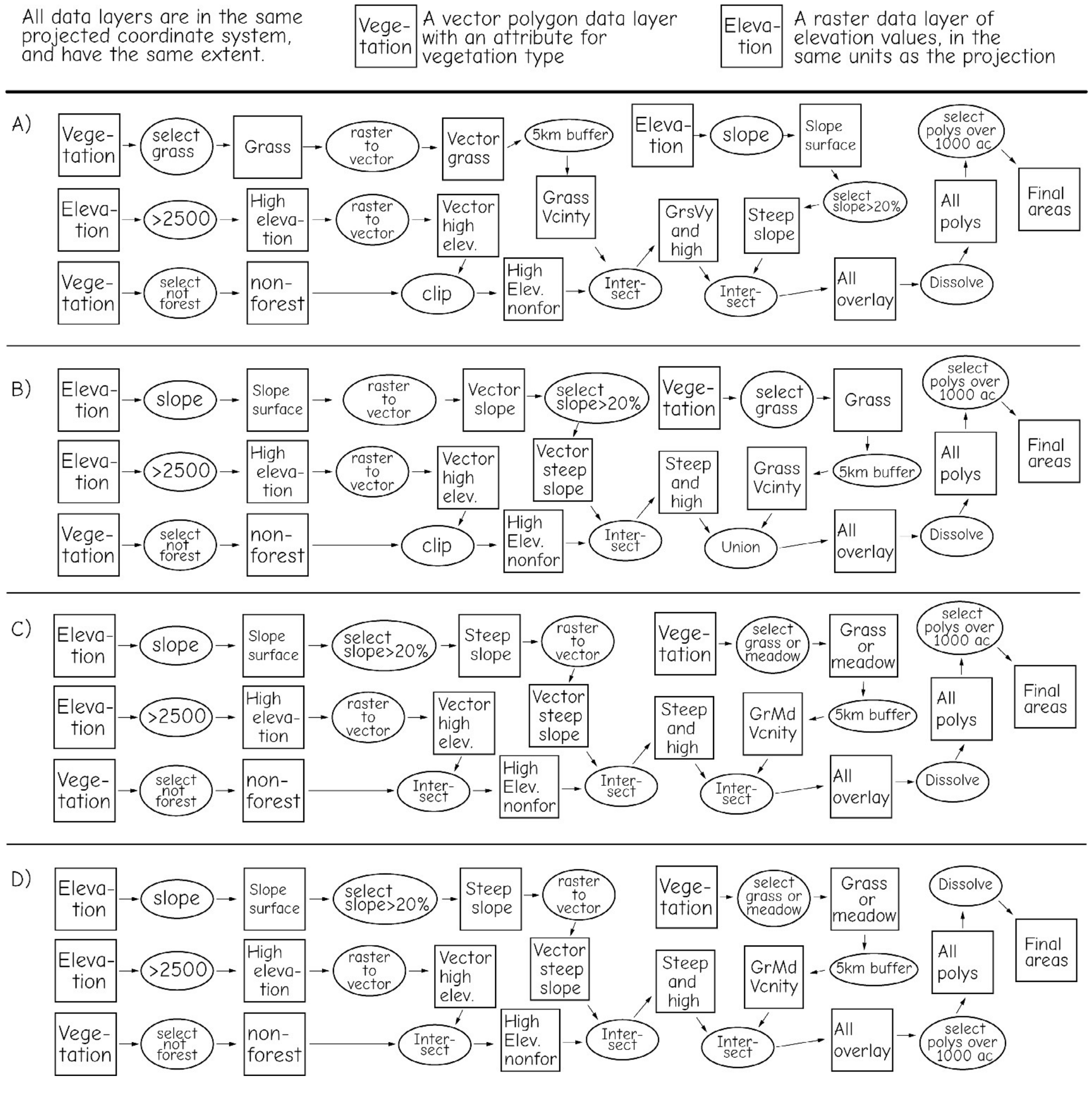
13.8 - The figure below depicts four flowcharts of cartographic models to find areas most suitable for new parks. Sites are preferred that meet all of the following criteria:
Within 0.5 miles of Census polygons with a density of more than 120 persons per square mile;
Greater than 0.5 miles from an existing park;
Within 0.5 miles of a school.

Select the flowchart that best approximates the proper analysis, given the described data. For each other flowchart, list at least one primary way it is inferior to the chosen method. Note that some intermediate steps are omitted for all flowcharts, so do not cite a step omitted in both the best and alternate flowcharts.



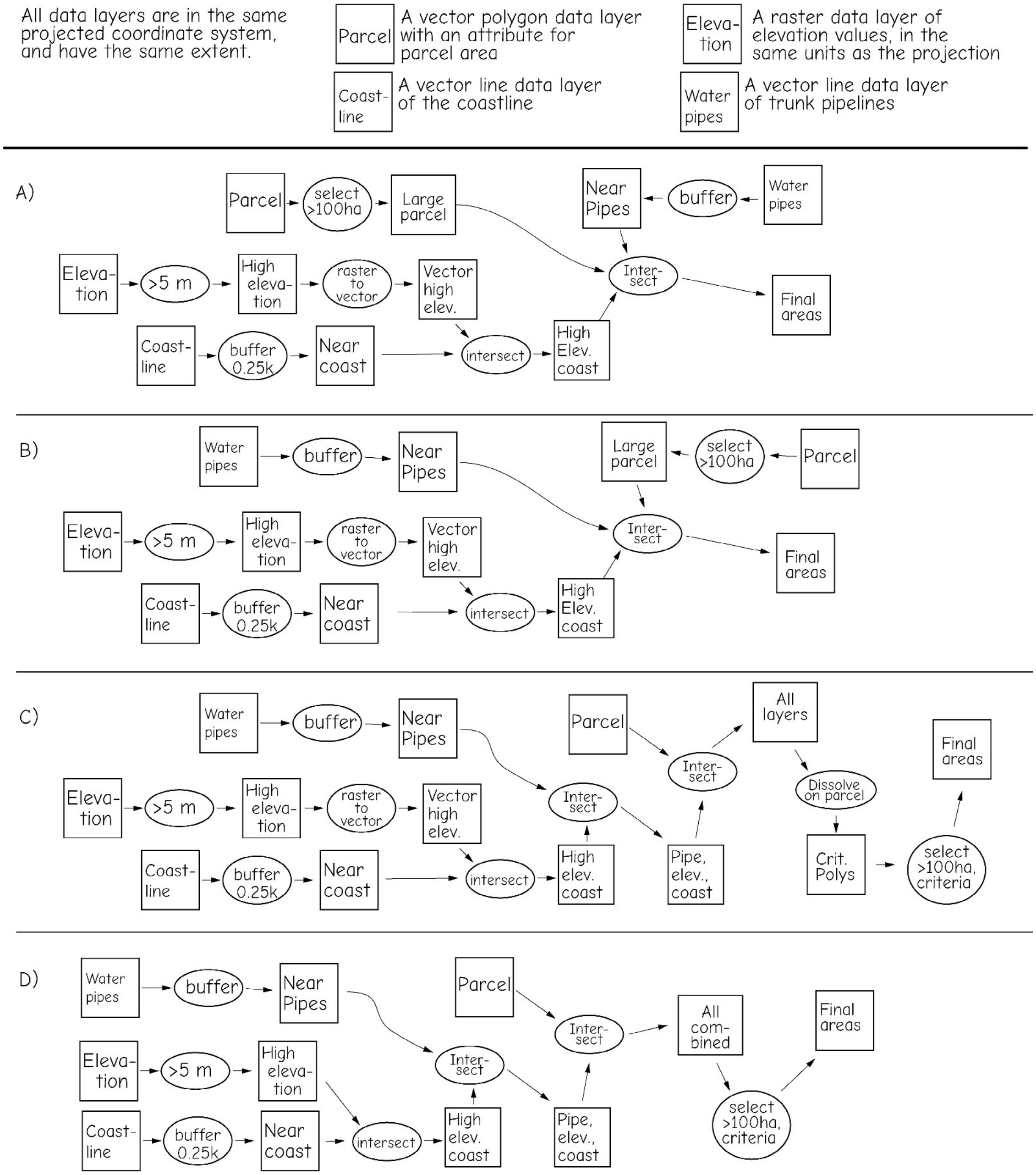
13.9 - The figure below depicts four flowcharts of cartographic models to find areas most suitable for wild sheep habitat. Sites are preferred that meet all of the following criteria:
Non-forest; Slope greater than 20%; Elevation above 2,500 m; All areas within 5 km of grassland or meadow; Each contiguous polygon larger than 1000 acres.

Select the flowchart that best approximates the proper analysis, given the described data. For each other flowchart, list at least one primary way it is inferior to the chosen method. Note that some intermediate steps are omitted for all flowcharts, so do not cite a step omitted in both the best and alternate flowcharts.



13.10 - The figure below depicts four flowcharts of cartographic models to find areas most suitable for a desalinization plant. Sites are preferred for which all parts meet all of the following criteria:
At an elevation of 5 m or greater;
Within 0.25 km of the coastline;
All areas within 0.5 km of trunk water pipe;
Within a single parcel;
Areas larger than 100 ha.

Select the flowchart that best approximates the proper analysis, given the described data. For each other flowchart, list at least one primary way it is more likely to lead to an error. Note that some small intermediate steps are omitted for all flowcharts, so do not cite a step omitted in both the best and alternate flowcharts.



14 Data Standards and Data Quality

Introduction

A *standard* is an established or sanctioned measure, form, or method. It is an agreed-upon way of doing something. Spatial data and analysis standards are important because of the range of organizations producing and using spatial data, and because these data are often transferred among organizations. Data standards facilitate a common understanding of the components of a spatial data set, how data were developed, and the utility and limitations of these data.

GIS practitioners use several types of standards. *Data standards* are used to format, assess, document, and deliver spatial data. *Interoperability standards* identify how spatial data are served between heterogeneous networks of software and hardware systems, for example, between wireless mobile devices and shared databases. *Analysis standards* ensure that the most appropriate methods are used and that the spatial analyses provide the best information possible. *Professional* or *certification standards* establish the education, knowledge, or experience of the GIS analyst, thereby improving the likelihood that the technology will be used appropriately.

We have progressed further in defining spatial data and interoperability standards than in defining analysis and professional standards. This is perhaps because GIS are used in such a wide range of disciplines. Urban planners, conservationists, civil and utility engineers, business people, and a number of other professions use GIS.

National and international standards organizations are important in defining and maintaining geospatial standards. The Federal Geographic Data Committee (FGDC) is the leading government organization in the United States in defining data standards. The FGDC focuses on the National Spatial Data Infrastructure (NSDI) in the United States, a set of resources to aid the creation and sharing of digital geographic data. Standards are developed through a set of processes, from proposals through drafts to a FGDC adopted standard. Standards may be modified through an update process. Currently, there are standards on methods (e.g., wetlands classification), content (Utilities Data Content Standard), metadata (data about data), and data transfer. Details are at www.fgdc.gov.

There are parallel initiatives in many countries, information on which can be found through the International Spatial Data Standards Commission. The Commission currently serves as a clearinghouse and gateway to national standards across the world.

The International Standards Organization (ISO) organizes international standards, and sponsors the ISO/TC211 standards (http://www.isotc211.org). These specify ways to store and represent spatial and

related information, services and data management, processing, transferring, and presenting information. The standards are organized as various projects, for example, standards for representing coordinates, testing standards, or for measuring data quality. Many standards are in active development, but inasmuch as these standards become stable, it will ease data and information transfer among different GIS software, among organizations, and through time.

Spatial data sharing happens across various software, computing platform, and physical or wireless networks. Interoperability standards are required because spatial data components are transferred across various systems and devices, often for use in real time. For example, coordinates and attribute data may be requested from an application on a smartphone across a cellular network, through a wired network to a server, the data accessed on a remote database, and served back through the networks to the field display.

The Open Geospatial Consortium (OGC) is an ad hoc, self-selected group of companies, research institutions, government bodies, and individuals dedicated to developing interoperability standards. Interoperation problems are identified, such as general difficulties in accessing time-varying spatial location data through a distributed wireless network, and standards for access proposed. These are reviewed, discussed, amended, and adopted.

Web mapping services (WMS) standards are an example of OGC initiatives. Web mapping services allow GIS software to access data across the internet as if they were stored on the local hard disk. A GIS program or utility “maps” the WMS to the local computer, meaning it may access the data with the same protocols as if it were stored locally, without downloading a permanent copy to store on the local hard disk.

Web services such as WMS are important for the future of *cloud-based computing*, where data, programs, and processing are seamlessly distributed on computers connected across the web. Cloud-based geospatial computing is inherently dependent on robust, well-defined interoperability standards such as those being developed by the OGC. Standards identify data formats and content, parts and naming, metadata, how connections are made and data are passed between programs across distributed networks, and error checking in transfer. Standards allow data to be combined across different organizations, with local storage and access form and protocols, and a standard way of serving up data to others through a service.

The *Indoor GML* is a newer OGC standard, under development to define data formats for interior building spatial data. Three-dimensional data for building interiors are useful to real estate, law enforcement, design, and construction applications. Various software vendors and research organizations have developed 3D formats, but data sharing is inhibited without standards. The OGC has developed such a data standard, with the participation of software, research, business, and government representatives.

It has proven more difficult to develop professional and analysis standards that are inclusive across all disciplines. Standard methods for one discipline may be inappropriate for another. For example, acceptable data collection methods for cadastral surveyors may be different than those for foresters. Cadastral surveys often require accuracies measured in centimeters (0.5 in) or less, while relatively sparse attribute information is recorded. Conversely, forest inventories may need only meter-level accuracies, but require a large set of attributes.

There has been recent progress on the development of a basic set of standards in the professional practice of GIS in the United States. Known as *competency models*, they define a set of skills considered essential for effective work in a field, and have been developed for a growing number of industries. All have a common foundation of basic personal and workplace competencies, with industry- and then occupancy-specific skills built on top.

The Geospatial Competency Model

The Geospatial Technology Competency Model (Figure 14-1) identifies a set of core and industry sector geospatial abilities. The Competency Model identifies examples of over 40 "Critical Work Functions" that geographic technology professionals are commonly expected to master and use in their careers, and the background knowledge on which these Critical Work Functions are based. The Geospatial Competency Model is based in part on the Geographic Information Science and Technology Body of Knowledge, first published by the Association of American Geographers in 2006. Critical work functions include operations in basic geodesy, data collection systems, data structures, GIS operation and programing, analytical methods, cartography, the place of geographic information science and technology in society, and organization and institutions. A set of higher-level requirements are noted for specific occupations.

We may reasonably expect this competency model to form the basis for professional certification, and to evolve into or form the basis of professional standards of knowledge in geospatial fields. One can envision professional or technical certification based on demonstrated knowledge in these areas, as in professional engineering exams, or perhaps in certification of completion of qualifying curricula. Currently there are no certification or testing mechanisms for this competency, but these may be developed in the future.

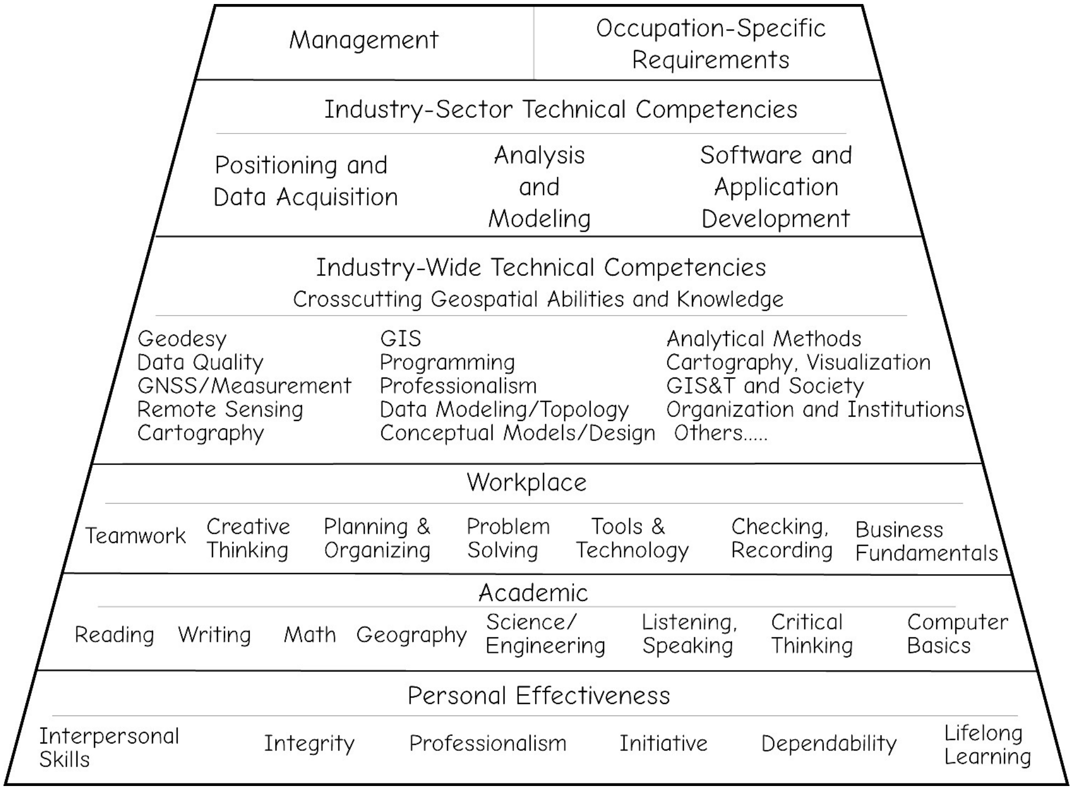


Figure 14-1: The Geospatial Competency Model developed by the U.S. Department of Labor. Please see: http://www.careeronestop.org/CompetencyModel/competency-models/geospatial-technology.aspx for a complete description of the model and components.

Spatial Data Standards

Spatial data standards can be defined as methods for structuring, describing, and delivering spatially referenced data. Spatial data standards may be categorized into four areas: media standards, format standards, accuracy standards, and documentation standards. All are important, although the last two are substantially more complex than the first two.

Media standards refer to the physical form in which data are transferred. They define specific formats for CD-ROM, magnetic tape, optical or solid state storage, or some proprietary drive or other media type. Standardized formats are specified by the International Standards Organization (ISO).

Format standards specify data file components and structures. A format standard establishes the number of files used to store a spatial data set, as well as the basic components contained in each file. The order, size, and range of values for the data element contained in each file are defined. Information such as spacing, variable type, and file encoding may be included.

Format standards aid in the practical task of transferring data between computer systems, either within or between organizations. Producers and users may not use the same hardware or GIS software. The interchange between different software systems is aided by general, standard forms in which data may be delivered.

Many government or vendor formats have become widely supported because data are commonly delivered using the formats. For example, the U.S. government supports the Spatial Data Transfer Standard (SDTS). This format specifies the logic, format, and encoding for raster, vector, and topological data transfer of spatial data. ESRI shapefiles (a cluster of files including .shp, .shx, and .dbf) are a commonly supported vector format, and many organizations transfer data using them. These proprietary formats are not truly standards because the formats may be changed by the vendors that created them. Until data formats are agreed to by a standardizing body, and there is some interpretation on how they are applied, hindering transparency and hence interoperability.

Spatial data accuracy standards document the quality of the positional and attribute values. Knowledge of data quality is crucial to the effective use of GIS, but we are often remiss in documenting spatial data quality. This is due in part to the cost of adequately estimating the errors in our spatial data sets. Field sampling is expensive. Data production is often pushed to available resources, and the documentation of data accuracy incomplete. Adherence to spatial data accuracy standards ensures we assess and communicate spatial data quality in a well-defined, established manner.

Documentation standards define how we describe spatial data. Data are derived from a set of original measurements taken by specific individuals or organizations at a specified time. Data may have been processed, and are stored in some format. Data documentation standards are an agreed-upon way of describing the source, development, and form of spatial data. When documentation standards are used, they ensure a complete description of the data origin, methods of development, accuracy, and delivery formats. Standard documentation allows the data steward to maintain the data, and these standards allow any potential user to assess the appropriateness of these data for an intended task.

Data quality standards add value to our data. There are many ways to describe data positional and attribute error. An incomplete description of spatial data quality may not allow a user to judge if the data are acceptable for an intended application. A data quality standard becomes familiar through use. We may know what levels of average error are likely to result in unacceptable data. The standard allows us to compare two data sets in light of this past experience.

Data Accuracy

An accurate observation reflects the true shape, location, or characteristics of the phenomena represented in a GIS. When the concept of accuracy is applied to spatial variables, it is a measure of how often or by how much our data values are in error. Accuracy may be reported as a frequency, for example, when we report that 20% of the land cover class labeled as cropland is actually perennial grasses. Alternatively, accuracy may be expressed as an average error magnitude; for example, light poles may be displaced on average by 12.4 m from their true locations.

Inadequacies in our spatial data model may cause spatial data error. When we use a raster data set with a fixed cell size, we have set a limit on our positional accuracy. The raster model assumes a homogeneous pixel. If more than one category or value for a variable is found in the pixel, then the attribute value may be in error. This generalization error may also occur in vector data sets. Any feature smaller than the minimum mapping unit may not be represented. Vector data sets may poorly represent gradual changes, so there can be increased attribute error near vector boundaries. Digital soils data are often provided in a vector data model, yet the boundaries between soil types are often not discrete, but change over a zone of a few to several meters.

Errors are often introduced during spatial data collection. Many positional data are currently collected using GNSS technologies. The spatial uncertainty in GNSS positions described in Chapter 5 is incorporated into the positional data. Feature locations derived from digitized maps or aerial photographs also contain positional errors due to optical, mechanical, and human deficiencies. Lenses, cameras, or scanners may distort images, positional errors may be introduced during registration, or errors may be part of the digitization process. Blunders, fatigue, or differences among operators in abilities or attitudes may result in positional uncertainty.

Spatial data accuracy may be degraded during laboratory processing or data reduction. Mis-copies during the transcription of field notes, errors during keyboard entry, or mistakes during data manipulation may alter coordinate values used to represent a spatial data feature. Improper representation in the computer may cause problems, such as rounding errors when multiplying large numbers.

Data may also be in error due to changes through time (Figure 14-2). The world is dynamic, while our representation in a spatial data set captures a snapshot at the time of data collection. Vegetation boundaries may be altered by fire, logging, construction, conversion to agriculture, or a host of other human or natural disturbances. Even in instances where positions are static, attributes may change through time. A two-lane gravel road may be paved or widened, causing attributes to be in error. Layers should have a recommended update interval that may vary by type. Elevation, geology, and soils may be updated rarely, and still maintain their accuracy. Vegetation, population, land use, or other factors change at faster rates, and should be updated more frequently if they are to remain accurate.

Documenting Spatial Data Accuracy

We must unambiguously identify true conditions if we are to document spatial data accuracy. For example, a road segment may be completely paved, or not. The data record for that road segment is accurate if it describes the surface correctly, and inaccurate if it does not. However, in many cases, the truth is not completely known. The locations for the above roads may be precisely surveyed using the latest carrier phase GNSS methods. Road centerlines and intersections may be known to the nearest 0.5 cm. While this is a very small error, this represents some ambiguity in what we deem to be the truth. Establishing the accuracy of a data set

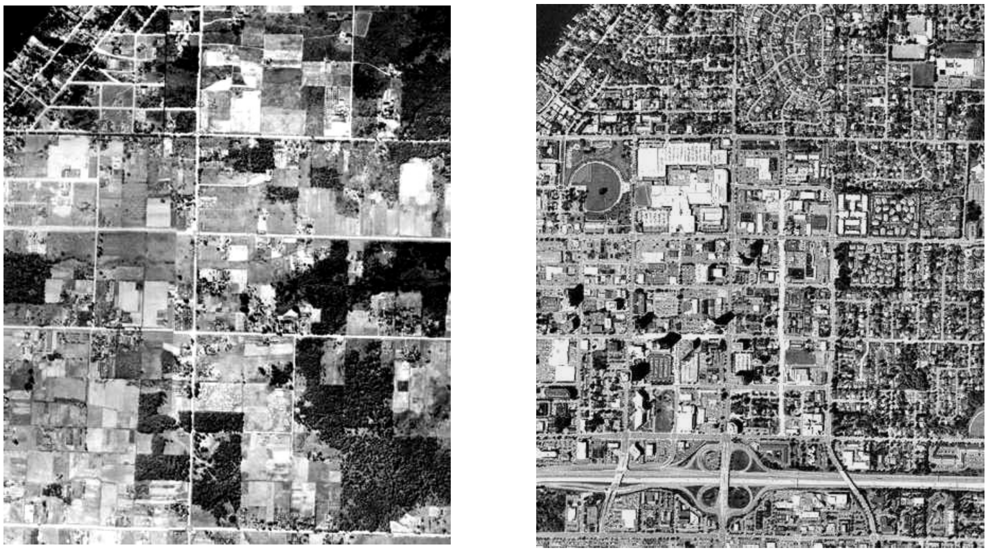


Figure 14-2: Spatial data may be in error because of the passage of time. Road maps based on 1936 photographs (left) from the city of Bellevue, Washington, are likely to be in error in 1997 (right) (courtesy Washington Department of Natural Resources).

requires we know the accuracy of our measure of truth.

In most cases, the truth is defined based on some independent, higher order measurements. In our roads example, we may desire that our data layer be accurate to 15 m or better. Gaged on this scale, the 0.5 cm accuracy from our carrier phase GNSS measurement may be considered true.

Accuracy is most reliably determined by a comparison of true values to the values represented in a spatial data set. This requires we collect data at an adequate set of sample locations. True values are collected at these sample locations. Corresponding values are collected for the digital spatial data. The true and data values are compared, errors calculated, and summary statistics generated.

The source for our truth, the sampling method, our method for calculating error, and the summary statistics we choose will depend on the type of spatial data that are to be evaluated. Positional data will be assessed using different methods than attribute data. Nominal attribute data (e.g., the type of land cover), will be assessed differently than a measurement recorded on a continuous range (e.g., purchase price of a parcel).

There are four primary ways we describe spatial data accuracy*: positional accuracy, attribute accuracy, logical consistency*, and *completeness* (Figure 14-3). These four components may be complemented with information on the *lineage* of a data set to define the accuracy and quality of a data set. These components are described in turn below.

Positional accuracy describes how close the locations of objects represented in a digital data set correspond to the true locations for the real-world entities. In practice, truth is determined from some higher-order positioning technology.

Attribute accuracy summarizes how different the attributes are from the true values. Attribute accuracies are usually reported as a mean error or quantile above a threshold error for attributes measured on interval/ratio scales, and as percentages or proportions accurate for ordinal or categorical attributes.

Logical consistency reflects the presence, absence, or frequency of inconsistent data. Tests for logical consistency often require comparisons among themes, for

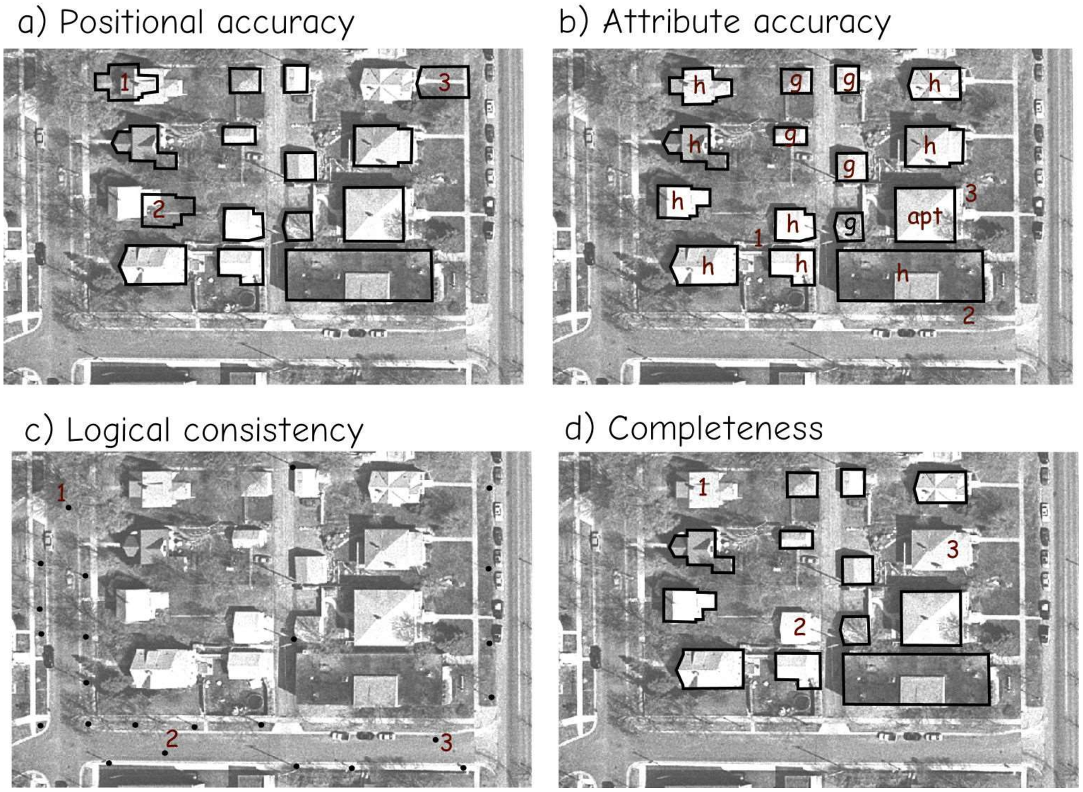


Figure 14-3: Examples of errors of various types. This figure shows digitized features (lines or points) overlain on a source aerial photograph. Errors are labeled with numbers. In a the houses labeled 1, 2, and 3 suffer from positional inaccuracy, while b demonstrates attribute accuracy in that garages at 1 are labeled as houses, and 2 and 3 show apartments and houses mis-labeled. Panel c shows that data may not be logically consistent, with light poles at locations 1, 2, and 3 in a street, and d shows that data may lack completeness, with houses 1, 2, and 3 not digitized, as shown by the missing outline.

example, all roads occur on dry land. This is different than positional accuracy in that both the road and the lake locations may contain positional error. However, these errors do not cause impossible or illogical juxtapositions. Logical consistency may also be applied to attributes, for example, wetland soils erroneously listed as suitable for construction, or lakes with zero depth.

Completeness describes how well the data set captures all the features. A buildings data layer may omit certain structures, and the frequency of these omissions reflects an incomplete data set.

Data sets may be incomplete because of generalizations during map production or digitizing. For example, a minimum mapping unit may be set at 2 ha when compiling a vegetation map. Isolated small pastures scattered through the forest may not be represented because they are only slightly larger than this minimum mapping unit, and erroneously they are not represented in the data layer.

Lineage describes the sources, methods, timing, and persons responsible for the development of a data set. Lineage helps establish bounds on the other measures of accuracy described above, because knowledge about certain primary data sources helps define the accuracy of a data set.

Positional Accuracy

Positional accuracy measures how close a database representation of an object is to the true value. Accurate positions have small errors. Small is defined subjectively, but may at least be quantified.

Precision refers to the consistency of a measurement method. Precision is usually defined in terms of how dispersed a set of

repeat measurements are from the average measurement. A precise measurement system provides tightly packed results. Precise digitizing means we may repeatedly place a point in the same location.

Accuracy and precision are often confused, but they are two different characteristics, both desirable, that may change independently. A set of measurements may be precise but inaccurate. Repeat measurements may be well clustered, meaning they are precise, but they may not be near the true value, meaning they are inaccurate. A *bias* may exist, defined as a systematic offset in coordinate values. A less precise process will result in a set of points that are more widely spread. However, their average error may be substantially less, therefore, the set is more accurate.

Figure 14-4 illustrates the difference between accuracy and precision. Four digitizing sessions are shown. The goal is to place several points at the center of the cloverleaf intersection in Figure 14-4. The upper left panel shows a digitizing process that is both accurate and precise. Points, shown as light-colored circles, are clustered tightly and accurately over the intended location.

The upper right panel of Figure 14-4 shows points that are precisely placed (tightly clustered), but not accurately located. This might be due to an equipment failure or some problem in registration; the operator may have made some blunder in photo registration and introduced a bias.

The lower left panel of Figure 14-4 shows points that are accurately but imprecisely digitized. The average location for these points is quite near the desired position, the center of the cloverleaf intersection, even though individual points are widely scattered. These points are not very close to the mean value and so precision is low, even though accuracy is high.

The panel at the lower right of Figure 14-4 shows points with positions that are

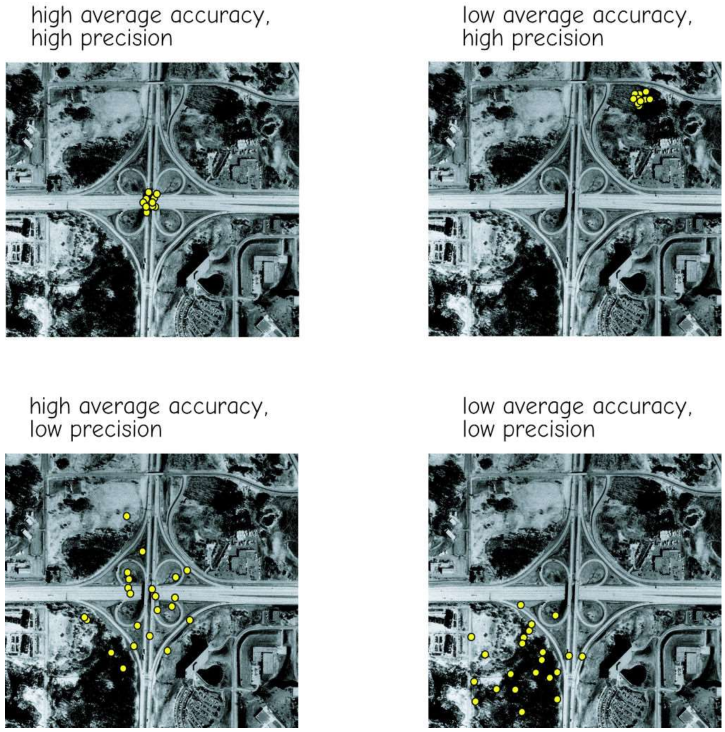


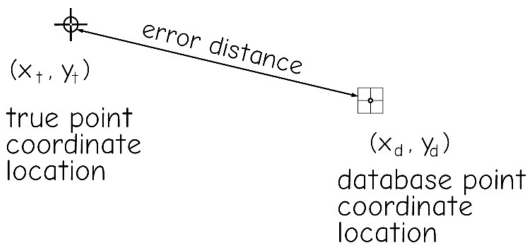
Figure 14-4: Accuracy and precision. Points (light yellow circles) are digitized to represent the center of the cloverleaf intersection. A group of points may be accurate, but not precise (lower left), or precise, but not accurate (upper right), or worst of all, not accurate, and not precise (lower left). We prefer a process that yields both accuracy and precision (upper left).

both imprecise and inaccurate. The mean value is not near the true location, nor are the values tightly clustered.

The thresholds that constitute high accuracy or precision are often subjectively defined. A duffer may consider as accurate any golf shot that lands on the green. This definition of accuracy may be based on thousands of previous attempts. For a professional golfer, anything farther than 2 m from the hole may be an inaccurate shot. In a similar fashion, the spatial accuracy sought by a land surveyor may be different than those of a federal land manager. Cadastral surveys require the utmost in accuracy because people tend to get upset when there is material permanent trespass, as when a neighbor builds a garage on their land. Lower accuracy is acceptable in other applications; for example, a statewide map defining vegetation type may be acceptable even though boundaries are off by tens of meters.

The mean error and an error frequency threshold are the statistics most often used to document positional data accuracy. Consider a set of wells represented as point features in a spatial data layer. Suppose that after we have digitized our well locations, we gain access to a GNSS system that effectively gives us the true coordinate locations for each well. We may then compare these well locations to the coordinate locations in our database. We begin by calculating the distance between our true and database coordinates for each well. This leaves us with a list of errors, one associated with each well location (Figure 14-5). Distance is measured using the Pythagorean formula with the true and database coordinates. Distances are always positive because of the form of the formula.

We may compute the mean error by summing the errors and dividing the sum by the number of observations. This gives us our average error, a useful statistic somewhere near the midpoint of our errors. We are often interested in the distribution of our errors, and so we also commonly use a frequency histogram to summarize our spatial error. The histogram is a graph of the number of error observations by a range of error values, for example, the number of error values between 0 and 1, between 1 and 2, between 2 and 3, and so on for all our observations. The graph will indicate the largest and smallest errors, and also give some indication of the mean and most common errors.



$$\text{error distance} = \sqrt{(x_t - x_d)^2 + (y_t - y_d)^2}$$

Figure 14-5: Positional errors are measured by the Pythagorean distance between a true and database coordinate for a location.

Examples of error frequency distributions for two different data sets are shown in Figure 14-6. Each plot shows the frequency of errors across a range of error distances. For example, the top graph shows that approximately 1 percent of errors have a value of near 4.5 m, and the mean error is near 13 m.

The mean error value does not indicate the distribution, or spread of the errors. Two data sets may have the same mean error but one may be inferior; the data set may have more large errors. The bottom graph in Figure 14-6 has the same mean error, 13 m, as the top graph. Note that the errors have a narrower distribution, meaning the errors are clumped closer to the mean than in the top graph, and there are fewer large errors. Although the mean error is the same, many would consider the data represented in the bottom graph of Figure 14-6 to be more accurate.

Because the mean statistic alone does not provide information on the distribution of positional errors, an error frequency threshold can be reported. An error fre-

quency threshold is a value above or below which a proportion of the error observations occurs. Figure 14-6 shows the 95% frequency threshold for two error distributions. The threshold is placed such that 95% of the errors are smaller than the threshold and 5% are larger than the threshold. The top graph shows a 95% frequency threshold of approximately 21.8 m. This indicates that approximately 95% of the positions tested from a sample of a spatial database are less than or equal to 21.8 m from the true locations. The bottom panel in Figure 14-6 has a 95% frequency threshold at 17.6 m. This means 5% of the errors in the second tested database are larger than 17.6 m from their true location. If we are concerned with the frequency of large errors, this may be a better summary statistic than the mean error.

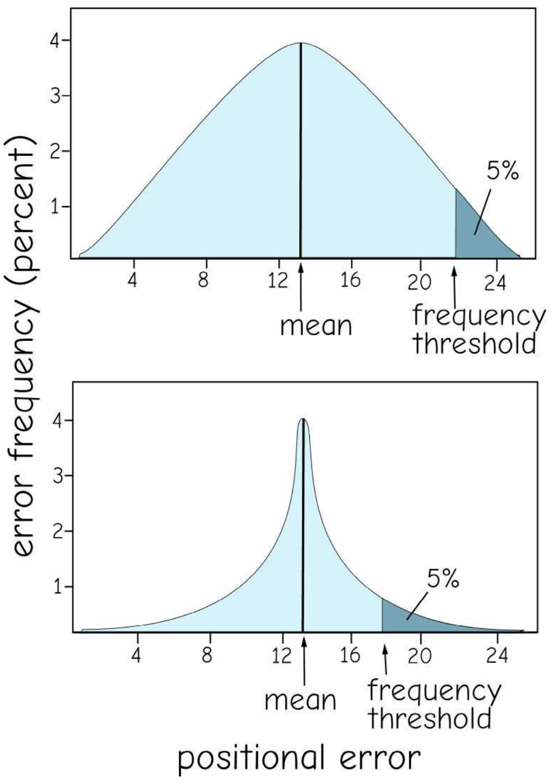


Figure 14-6: Mean error and frequency thresholds are often used to report positional error. The lightly shaded gray area represents 95% of the data.

A Standard Method for Measuring Positional Accuracy

The Federal Geographic Data Committee of the United States (FGDC) has described a standard for measuring and reporting positional error. It is known as the National Standard for Spatial Data Accuracy (NSSDA). The NSSDA specifies the number and distribution of sample points when performing an accuracy assessment, and prescribes the statistical methods used to summarize and report positional error. Separate methods are described for horizontal (X and Y) accuracy assessment and vertical (Z) accuracy assessment, although the methods differ primarily in the calculation of summary accuracy statistics. There are five steps in applying the NSSDA:

- Identify a set of test points from the digital data set under scrutiny;
- Identify a data set or method from which “true” values will be determined;
- Collect positional measurements from the test points as they are recorded in the test and “true” data sets;
- Calculate the positional error for each test point and summarize the positional accuracy for the test data set in a standard accuracy statistic;
- Record the accuracy statistic in a standardized form. Also include a description of the sample number, true data set, the accuracy of the true data set, and the methods used to develop and assess the accuracy of the true data set.

Test points must be clearly identifiable in both the test data set and in the truth data set. Points that are precisely, unambiguously defined are best. For example, we may wish to document the accuracy of roads data compiled from medium-scale sources and represented by a single line in a digital layer. Right-angle road intersections are preferred over other features because the positions represented in the database may be precisely determined. The coordinates for the precise center of the road intersections may also be

determined from a higher-accuracy data set, for example, from digital orthophotographs or field surveys. Other road features are less appropriate for test points, including road intersections at obtuse angles or acute curves, because there may be substantial uncertainty when matching the data layer to true coordinates.

The source of the true coordinate position should match our minimum accuracy specification, or at least an order of magnitude more accurate than the errors. GNSS are a common source of truth, as the accuracy may be set by collection equipment and methods, but any source of truth that matches our requirements is acceptable.

Figure 14-7 shows an example set of test points for road data layer, and an image backdrop. Prior knowledge leads us to expect average errors in excess of 6 m. In this example, we have selected high resolution GNSS as our true data source. We know these photographs have a positional error of less than 15 cm, on average, from metadata. These images were selected because they meet our accuracy requirements and are available for the entire work area.

The display of road locations on top of the images shows there are substantial differences in true positions of features and their representations in the roads data layer. Any right-angle intersection is a prospective test point.

The inset in the lower left of Figure 14-7 shows the true point locations relative to the road intersections. Road centerlines were digitized. These true locations would be identified on the images, perhaps by point-

Figure 14-7: A roads data layer displayed over a georeferenced image. Test point locations are shown as filled circles. The test point true coordinate values should be from a high accuracy source, e.g., centimeter level GNSS surveys. The data coordinate values are extracted from the roads layer. Differences in these locations would be used to estimate the positional accuracy of the roads data layer.

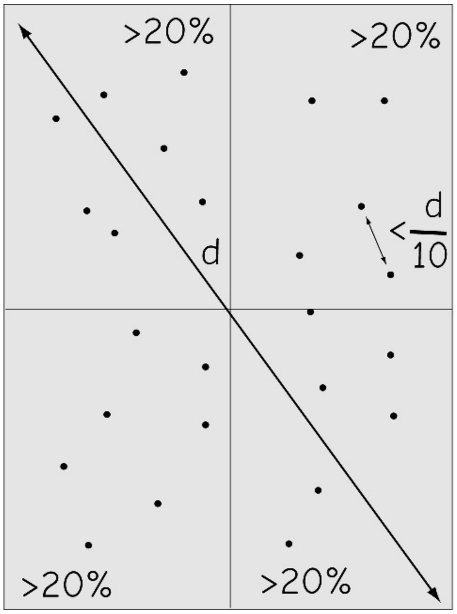


Figure 14-8: Recommended sampling for spatial data accuracy. Samples should be well distributed and well spaced (adapted from LMIC, 1999).

ing a cursor at a georeferenced image displayed on a computer monitor. The data coordinates would then be extracted for the corresponding road intersection, and these two coordinate pairs, the true X, Y and data X, Y, would be one test point used in accuracy calculations.

The NSSDA specifies between 20 and 30 well-distributed test points (Figure 14-8). Test points should be distributed as evenly as possible throughout the data layer to be tested. Each quadrant of the tested data layer should contain at least 20% of the test points, and test points should be spaced no closer than one-tenth the longest spanning distance for the tested data layer (d, in Figure 14-8).

Accuracy Calculations

The calculation of point accuracies and summary statistics are the next steps in accuracy assessment. First, the coordinates of both the true and data layer positions for a feature are recorded. These coordinates are used to calculate a positional difference, known as a positional error, based on the distance between the true coordinates and the data layer coordinates (Figure 14-5). The true coordinates fall in a different location than the coordinates derived from the data layer. Each test point yields an error distance e, shown in Figure 14-5 and defined by the equation:

$$e = \sqrt{(x_t - x_d)^2 + (y_t - y_d)^2} \quad (14.1)$$

where x_t, y_t are true coordinates and x_d, y_d are the data layer coordinates for a point.

The squared error differences are then calculated, and the sum, average, and root mean square error (RMSE) statistics determined for the data set. As previously defined in this book, the RMSE is:

$$RMSE = \sqrt{\frac{e_1^2 + e_2^2 + \ldots e_n^2}{n}} \quad (14.2)$$

where e is defined as in equation 14.1, and n is the number of test points used.

The RMSE is not the same as the average distance error, nor a "typical" distance error. The RMSE is a statistic that is useful in determining probability thresholds for error. The RMSE is related to the statistical variance of the positional error. If we assume the x and y errors follow a bell-shaped Gaussian curve commonly observed when sampling, then the RMSE tells us something about the distribution of distance errors. We can use knowledge about the RMSE that we get from our sample to determine what is the likelihood of a large or small error. A large RMSE means the errors are widely spread, and a small RMSE means the errors are packed tightly around the mean value.

Statistical theory allows us to establish fixed numbers that identify error thresholds. Because we have two variables, X and Y, if we make appropriate assumptions, we can fix an error threshold at a given value. An error threshold is commonly set for 95%. When we fix a 95% error threshold, this means we identify the specific number such that 95% of our errors are expected to be less

than or equal to the threshold. Statistical theory tells us that when we multiply the RMSE by the number 1.7308 and assume a Gaussian normal distribution, we obtain the 95% threshold. A thorough treatment of the statistical foundation may be found in the references listed at the end of this chapter.

Accuracy calculations may be summarized in a standard table, shown in Table 14-1. The example shows a positional accuracy assessment based on a set of 22 points. Data for each point are organized in rows. The true and data layer coordinates are listed, as well as the difference and difference squared for both the x and y coordinate directions. The squared differences are summed, averaged, and the RMSE calculated, as shown in the summary boxes in the lower right portion of Table 14-1. The RMSE is multiplied by 1.7308 to estimate the 95% accuracy level, listed as the NSSDA. Ninety-five percent of the time, the true horizontal errors are expected to be less than the estimated accuracy level of 12.9 m listed in Table 14-1.

Table 14-1: An accuracy assessment summary table.

ID	x (true)	x (data)	x difference	(x difference)^2	y (true)	y (data)	y difference	(y difference)^2	sum x diff^2 + y diff^2
1	12	10	2	4	288	292	-4	16	20
2	18	22	-4	16	234	228	6	36	52
3	7	12	-5	25	265	266	-1	1	26
4	34	34	0	0	243	240	3	9	9
5	15	19	-4	16	291	287	4	16	32
6	33	24	9	81	211	215	-4	16	97
7	28	29	-1	1	267	271	-4	16	17
8	7	12	-5	25	273	268	5	25	50
9	45	44	1	1	245	244	1	1	2
10	110	99	11	121	221	225	-4	16	137
11	54	65	-11	121	212	208	4	16	137
12	87	93	-6	36	284	278	6	36	72
13	23	22	1	1	261	259	2	4	5
14	19	24	-5	25	230	235	-5	25	50
15	76	80	-4	16	255	260	-5	25	41
16	97	108	-11	121	201	204	-3	9	130
17	38	43	-5	25	290	288	2	4	29
18	65	72	-7	49	277	282	-5	25	74
19	85	78	7	49	205	201	4	16	65
20	39	44	-5	25	282	278	4	16	41
21	94	90	4	16	246	251	-5	25	41
22	64	56	8	64	233	227	6	36	100
								Sum	1227
								Average	55.8
								RMSE	7.5
								NSSDA	12.9

Errors in Linear or Area Features

The NSSDA as described above treats only the accuracies of point locations. It is based on a probabilistic view of point locations. We are not sure where each point is; however, we can specify an error distance r for a set of features. A circle of radius r centered on a point feature in our spatial data layer will include the true point location 95% of the time. Unfortunately, there are no established standards for describing the accuracy or error of linear or area features.

In some instances, we may assume the well-defined point features described in our accuracy test above may also represent the accuracy for nodes and vertices of lines in a data layer. Nodes or vertices may be used as test points, provided they are well defined and the true coordinates are known. However, the errors at intervening locations are not known, for example, midway along a line segment between two vertices. The error along a straight line segment may be at most equal to the largest error observed at the ends of the line segments (Figure 14-9). If the data line segment is parallel to the true line segment, then the errors are uniform along the full length of the segment. Vertices that result in converging or crossing lines will lead to midpoint errors less than the larger of the two errors at the endpoints (Figure 14-9). These observations are not true if a straight line segment is used to approximate a substantially curved line. However, if the line segments are sufficiently short (e.g., the interval along the line is small relative to the radius of a curve in the line), and the positional errors are distributed evenly on both sides of the line segments, then the NSSDA methods described above will provide an approximate upper limit on the linear error.

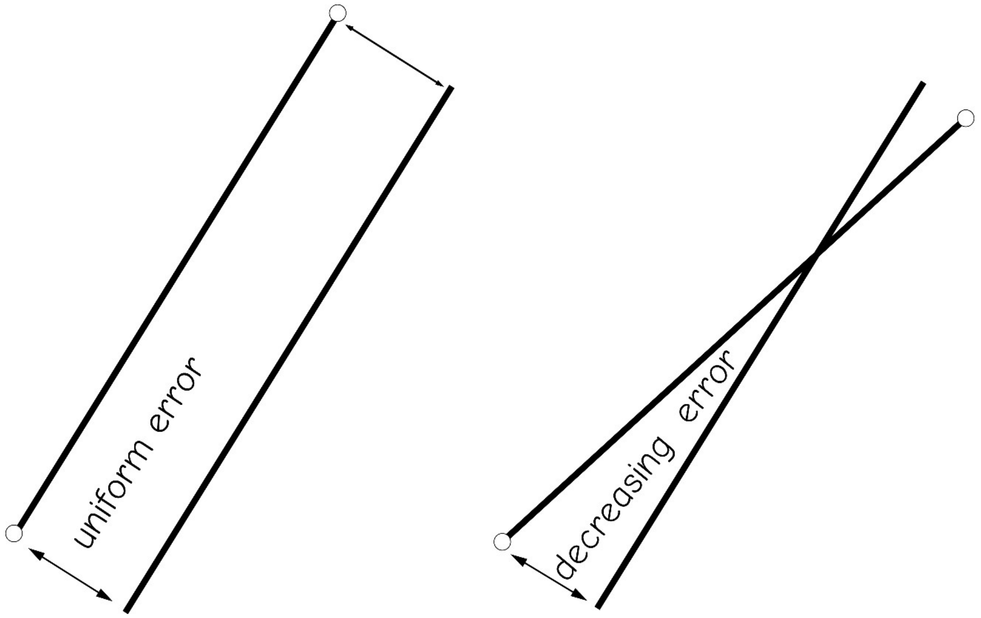


Figure 14-9: Errors for straight line segments are either the same (left) or less than (right) the maximum error observed at the end points. If nodes or vertices are sufficiently close such that the true line segments are approximately linear, then error assessments on nodes or vertices may provide an approximation of the line positional errors.

Attribute Accuracy

Unlike positional accuracy, there is no national standard for measuring and reporting attribute accuracy. Accuracy for continuous variables may be calculated in an analogous manner to positional accuracy. Accuracy for each observation is defined as the difference between the true and database values. A set of test data points may be identified, the true attribute value determined for each of those test data points, the difference calculated for each test point, and the accuracy summarized.

The accuracy of categorical attribute data may be summarized using an *error table* and associated accuracy statistics. Points can be classified as correct, that is, the categorical variable matches the true category for a feature, or they may be incorrect. Incorrect observations occur when the true and layer category values are different. Error tables, also known as error matrices, confusion matrices, and accuracy tables, are a standard method of reporting error in classified remotely sensed imagery. They have more rarely been used for categorical attribute accuracy assessment.

An error table summarizes a two-way classification for a set of test points (Figure 14-10). A categorical variable will have a fixed number of categories. These categories are listed across the columns and along the rows of the error table. Each test feature is tallied in the error table. The true category and the value in the data layer are known for each test feature. The test feature is tallied in the error table based on these values. The true values are entered via the appropriate column and the data layer values are entered

true value

data layer attribute value

	wheat	corn	soy	alfalfa	grass	fallow	
wheat	14	4			4		22
corn	2	12		1	3		18
soy	1		18	2			21
alfalfa		3	2	16	1		22
grass	3	1		1	12		17
fallow						20	20
	20	20	20	20	20	20	$\Sigma = 92$

$$\text{overall accuracy} = \frac{\text{sum of diagonal}}{\text{total number of samples}} = 92/120 = 76.7\%$$

Figure 14-10: An error table succinctly summarizes the attribute accuracy for categorical variables.

via the appropriate row. The table is square, because there is the same number of categories in both the rows and columns. Correctly classified features are tallied on the diagonal – the true value and data layer value are identical, so they are noted at the intersection of the categories. Incorrectly assigned category values fall off the diagonal.

Error tables summarize the main characteristics of confusion among categories. The diagonal elements contain the test features that are correctly categorized. The diagonal sum is the total number correct. The proportion correct is the total number correct divided by the total number tested. The percent correct can be obtained by multiplying the proportion correct by 100.

Per category accuracy may be extracted from the error table. Two types of accuracy may be calculated, a *user's accuracy* and a *producer's accuracy*. The user relies on the data layer to determine the category for a feature. The user is most often interested in how often a feature is mislabeled for each category. In effect, the user wants to know how many features that are classified as a category (the row total) are truly from that category (the diagonal element for that row). Thus, the user's accuracy is defined as the number of correctly assigned features (the diagonal element) divided by the row total for the category. The producer, on the other hand, knows the true identity of each feature and is often interested in how often these features are assigned to the correct category. The producer's accuracy is defined as the diagonal element divided by the column total.

Error Propagation in Spatial Analysis

While we have discussed methods for assessing positional and attribute accuracy, we have not described how we determine the effects of input errors on the accuracy of spatial operations. Clearly, input error affects output values in most calculations. A large elevation error in DEM cells will likely cause errors in slope values. If slope is then combined with other features from other data layers, these errors may in turn propagate through the analysis. How do we assess the propagation of errors and their impacts on spatial analysis?

There are currently no widely applied, general methods for assessing the effect of positional errors on spatial models. Research is currently directed at several promising avenues; however, the range of variables and conditions involved has confounded the development of general methods for assessing the impacts of purely positional errors on spatial models.

Several approaches have been developed to estimate the impacts of attribute errors on spatial models. One approach involves assessing errors in the final result irrespective of errors in the original data. For example, we may develop a cartographic model to estimate deer density in a suburban environment. The model may depend on the density of housing, forest location, type, and extent, the location of wetlands, and road location and traffic volumes. Each of these data sources may contain positional and attribute errors.

Questions may arise regarding how these errors in our input data affect the model predictions for deer density. Rather than trying to identify how errors in the input propagate through to affect the final model results, we may opt to perform an error assessment of our final output. We would perform a field survey of deer density and compare the values predicted by the model with the values observed in the field. For example, we might subdivide the study area into mutually exclusive census areas. Deer might be counted in each census area and the density calculated. We have replicated values from each census area, so we may calculate a mean and a variance, and the difference between modeled and observed values might be compared relative to the natural variation we observe among different census areas. We could also survey an area through time, for example, on successive days, months, or years, and compare the dif-

ference between the model and observed values for each sample time.

It may not be possible or desirable to wait to assess accuracy until after completing a spatial analysis. Input data for a specific spatial analysis may be expensive to collect. We may not wish to develop the data and a spatial model if errors in the input preclude a useful output. After model application, we may wish to identify the source of errors in our final predictions. Improvements in one or two data layers may substantially improve the quality of our predictions; for example, better data on forest cover may increase the accuracy of our deer density predictions.

Error propagation in spatial models is often investigated with repeated model runs. We may employ some sort of repeat simulation model that adds error to data layers and records the impacts on model accuracy. These simulation models often employ a standard form known as a *Monte Carlo simulation*. The Monte Carlo method assumes each input spatial value is derived from a population of values. For example, landcover may range over a set of values for each cell. Further, model coefficients may also be altered over a range. In a cartographic model, the weights are allowed to range over a specified interval when layers are combined.

A Monte Carlo simulation controls how these input data or model parameters are allowed to vary. Typically, a random normal distribution is assumed for continuous input values. If all variables save one are held constant, and several model runs performed on different, random selections of the variable, we may get an indication of how a variable affects the model output. We may find that the spatial model is insensitive to large changes in most of our input data values, but sensitive to small changes in a few. For example, predicted deer density may not change much even when landcover varies over a wide range of values, but may depend heavily on housing density. However, we may also find a set of input data, or a range of input data or coefficients, that substantially control model output.

A Monte Carlo or similar simulation is a computationally intensive technique. Thousands of model runs are often required over each of the component units of the spatial domain. The computational burden increases as the models become more involved, and as the number of spatial units increases. However, it is often the only practical way with which to assess the impacts of uncertainties on spatial analyses, uncertainties both in the input data and the parameters and methods in combining them.

Summary

Data standards, data accuracy assessment, and data documentation are among the most important activities in GIS. We cannot effectively use spatial data if we do not know its quality, and the efficient distribution of spatial data depends on a common understanding of data content.

Data may be inaccurate due to several causes. Data may be out of date, collected using improper methods or equipment, or collected by unskilled or inattentive persons.

Accuracy is a measure of error, a difference between a true and represented value. Inaccuracies may be reported using many methods, including a mean value, a frequency distribution, or a threshold value. An accuracy assessment or measurement applies only to a specific data set and time.

Accuracy should be recognized as distinct from precision. Precision is a measure of the repeatability of a process. Imprecise data collection often leads to poor accuracy.

Standards have been developed for assessing positional accuracy. Accuracy assessment and reporting depend on sampling. A set of features is visited in the field, and the true values collected. These true values are then compared to corresponding values stored in a data layer, and the differences between true and database values quantified. An adequate number of well-distributed

samples should be collected. Standard worksheets and statistics have been developed.

Data documentation standards have been developed in the United States. These standards, developed by the Federal Geographic Data Committee, are known as the Content Standard for Digital Geospatial Metadata. This standard identifies specific information that is required to fully describe a spatial data set.

Suggested Reading

Arbia, G., Griffith, D., Haining, R. (1999). Error propagation and modeling in raster GIS: overlay operations. *International Journal of Geographical Information Science*, 12:145–167.

Balazinska, M., Deshpande, A., Franklin, M.J., Gibbons, P.B., Gray, J., Hansen, M., Liebhold, M., Nath, S., Szalay, A., Tao, V. (2007). Data management in the Worldwide Sensor Web. *Pervasive*, 6:30–40.

Blakemore, M. (1984). Generalization and error in spatial data bases. *Cartographica*, 21:131–139.

Bolstad, P., Gessler, P., Lillesand, T. (1990). Positional uncertainty in manually digitized map data. *International Journal of Geographical Information Systems*, 4:399–412.

Chong, A.K. (1999). A technique for spatial sampling and error reporting for image map bases. *Photogrammetric Engineering & Remote Sensing*, 65:1195–1198.

Comber, A.J., Fisher, P.F., Harvey, F., Gahegan, M., Wadsworth, R. (2006). Using metadata to link uncertainty and data quality assessments. *Progress in Spatial Data Handling*, 6:279–292.

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A.T., Plewe, B., Wentz, W. (2006). *Geographic Information Science and Technology Body of Knowledge*. Association of American Geographers.

DiBiase, D., Corbin, T., Fox, T., Francica, J., Green, K., Jackson, J., Jeffress, G., Jones, B., Jones, B., Mennis, J., Schuckman, K., Smith, C., Van Sickle, J. (2010). The new Geospatial Technology Competency Model: Bringing workforce needs into focus. *URISA Journal*, 22:55-72.

Dunn, R., Harrison, A.R., White, J.C. (1990). Positional accuracy and measurement error in digital databases on land use: an empirical study. *International Journal of Geographical Information Systems*, 4:385–398.

Fisher, P. (1991). Modelling soil map unit inclusions by Monte Carlo simulation. *International Journal of Geographical Information Systems*, 5:193–208.

Goodchild, M.F. and Gopal, S. (1989). *The Accuracy of Spatial Databases*. London: Taylor and Francis.

Guptill, S.C. Morrison, J.L.(Eds.) (1995). *Elements of Spatial Data Quality*. New York: Elsevier.

Harmel, R.D., Smith, D.R., King, K.W., Slade, R.M. (2009). Estimating storm discharge and water quality data uncertainty: A software tool for monitoring and modeling applications. *Environmental Modeling and Software*, 24:832–842.

Heuvelink, G. (1999). *Error Propagation in Environmental Modeling with GIS*. London: Taylor and Francis.

Heuvelink, G., Brown, J.D., van Loon, E.E. (2007). A probabilistic framework for representing and simulating uncertain environmental variables. *International Journal of Geographical Information Science*, 21: 497–513.

Hunsacker, C.T., Goodchild, M.F., Friedl, M.A., Case, T.J. (2001). *Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications*. New York: Springer–Verlag.

Jones, R.R., McCaffrey, K.J.W., Wilson, R.W., Holdsworth, R.E. (2004). Digital field data acquisition: towards increased quantification of uncertainty during geological mapping. *Geological Society of London Special Publications*, 239:43–56.

Kassenberg, D., De Jong, K. (2005). Dynamic environmental modeling in GIS: 2. Modeling error propagation. *International Journal of Geographical Information Science*, 19:623–637.

LMIC (1999). *Positional Accuracy Handbook: Using the National Standard for Spatial Data Accuracy to measure and report geographic data quality*. Minnesota Planning, St. Paul.

Lodwick, W.A., Monson, W., Svoboda, L. (1990). Attribute error and sensitivity analysis of map operations in geographical information systems. *International Journal of Geographical Information Systems*, 4:413–427.

Lowell, K., Jaton, A. (1999). *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*. Chelsea: Ann Arbor Press.

Thapa, K. and Bossler, J. (1992). Accuracy of spatial data used in geographic information systems, *Photogrammetric Engineering and Remote Sensing*, 58:841–858.

Walsh, S.J., Lightfoot, D.R., Butler, D.R. (1987). Recognition and assessment of error in geographic information systems, *Photogrammetric Engineering and Remote Sensing*, 53:1423–1430.

Study Questions

14.1 - Why are standards so important in spatial data?

14.2 - Can you describe processes or activities that are greatly helped by the existence of standards?

14.3 - What are the differences between accuracy and precision?

14.4 - How do mean and frequency thresholds differ in the way they report positional error?

14.5 - What are some of the primary causes of positional error in spatial data?

14.6 - Describe each of the following concepts with reference to documenting spatial data accuracy: positional accuracy, attribute accuracy, logical consistency, and completeness.

14.7 - What is the NSSDA, and how does it help us measure positional accuracy?

14.8 - What are the basic steps in applying the NSSDA?

14.9 - What are the constraints on the distribution of sample points under the NSSDA, and why are these constraints specified?

14.10 - What are good candidate sources for test points in assessing the accuracy of a spatial data layer?

14.11 - How are errors in nominal attribute data often reported?

14.12 - What are metadata, and why are they important?

15 New Developments in GIS

Introduction

As every economist, weather forecaster, or politician knows, predicting the future is fraught with peril. Near-term predictions may be safe; if times are good now, they will probably be good next month. However, the farther one reaches into the future, the more likely they will be wrong. This chapter describes technologies that may become widespread. It discusses future trends, with the expectation that many of these speculations will prove inaccurate.

Many changes in GIS are based on advances in computers and other electronic hardware. Computers are becoming smaller and less expensive. This is true for both general-purpose machines and for specialized computers, such as ruggedized, portable tablet computers. The wizards of semiconductors continue to dream up and then produce impossibly clever devices. Given current trends, we should not be surprised in the future if a pea-sized device holds all the published works of humankind. Computers may gain personalities, recognize us as individuals, respond entirely to voice commands, and routinely conjure three-dimensional images that float in space before our eyes. These and other developments will alter how we manipulate spatial data.

Changes in GIS will also be due to the growing ubiquity of high speed, wireless and wired connections. If our data are always available, we will interact with them differently. We can more easily see how things should be in the field, and compare them to how they are, for example, a wiring diagram for a roadside telephone interchange panel, or a building site plan vs. stakeout. An agricultural field's fertilization history, a water main's flange size, or bridge's inspection records may all be available at any time, anyplace, streamlining maintenance and management.

Change is also due to increased sophistication in GIS software and users, and increased familiarity and standardization. Change will be driven by new algorithms or methods, for example, improved data compression techniques that speed the retrieval and improve the quality of digital images. Specialized software packages may be crafted that turn a multiday, technically complicated operation into a few mouse clicks. These new tools will be introduced as GIS technologies continue to evolve and will change the way we gather and analyze spatial data.

GNSS

Three trends will dominate GNSS innovation over the next decade: multi-constellation and multi-signal GNSS receivers, miniaturization, and system integration. Multi-GNSS receivers will continue to take advantage of distinct satellite constellations. GNSS has been unable to provide 10 cm (subfoot) position accuracies in thick forests, deep valleys, or city centers. Dual GPS/GLONASS systems already exist, and systems that simultaneously support the Galileo and Chinese Compass system will be developed, further increasing accuracy, availability, and reliability. Receivers will commonly have hundreds of channels, and track tens of satellites even when under heavy forest canopies, in canyons, and among tall buildings, bringing real-time precise positioning to everyone.

Dual channel GNSS chips will provide greatly improved accuracy at substantially reduced prices (Figure 15-1). Most inexpensive GNSS receivers have to date tracked a single frequency, e.g., L1 in the U.S. GPS system. Measurements in a second frequency allows reduction of ionospheric and atmospheric transit errors, leaving smaller system and clock errors. With suitable processing, the averaging/error prediction, centimeter-level accuracies will be available in real time, for tens of dollars in hardware cost.

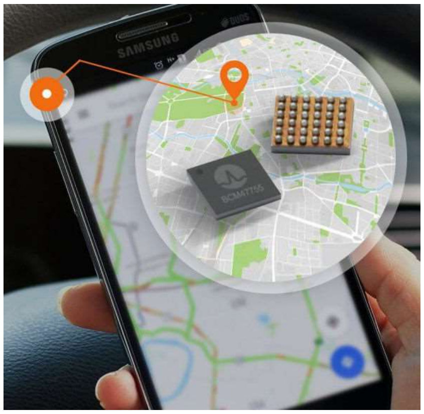


Figure 15-1: Dual frequency chips will provide cm-level accuracies to a broad array of devices (courtesy Broadcom).

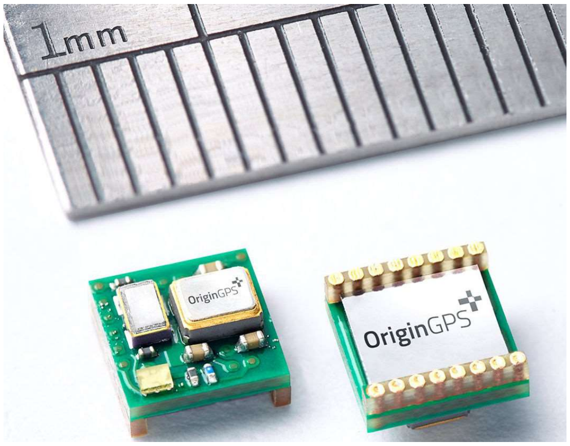


Figure 15-2: A miniaturized GNSS unit that may be embedded in a range of electronic devices.

GNSS receivers will cost less, shrink in size and weight, and increase accuracy for some time to come, and these improvements will spur even more widespread adoption of this technology (Figure 15-2). Microelectronic miniaturization is helping shape the GNSS market. As GNSS use grows and manufacturing methods improve, single chip GNSS systems have emerged, and these chips are decreasing in size. GNSS chips smaller than a postage stamp are available, including some that may be integrated into common electronic devices. Many vendors are well on a path to system integration, and it will become more common to embed the antenna, receiver, supporting electronics, power supply, and differential correction radio receivers in a single piece of equipment. Some of these integrated systems are smaller than most GNSS antennas of a decade ago, and systems will continue to shrink. A button-sized GNSS is not far off.

As receivers shrink in size and cost, it becomes practical to collect positional information on smaller individual objects. While GNSS is unlikely to help you find your keys, small GNSS receivers will collect spatial data for smaller objects. For example, a few years ago it was uneconomical to track objects smaller than a cargo ship. Now

trucks or containers are routinely followed. In the near future, it may be common to track individual packages.

GNSS miniaturization means we will directly collect much more data in the field than in times past. A city engineer may study traffic patterns by placing special-purpose GNSS receivers into autos. How long does the average commute take? How much of the time is spent sitting at stop signs or lights, and where is the congestion most prevalent? How is traffic affected by weather conditions? Analyses of traffic networks will become substantially easier with small-unit GNSS. Disposable GNSS receivers may be pasted, decal-like, on windshields by the thousands, to transmit their data back to a traffic management center.

Ubiquitous, inexpensive, or free differential correction signals are substantially improving the accuracy commonly achieved with GNSS. Many states in the United States, and national governments abroad will establish more complete coverage. Virtual Reference Station (VRS) networks promise to allow submeter and even near-centimeter level positioning in real time. Commercial solutions will be further developed and made less expensive.

GNSS systems will add functions, including the ability to take photos or videos and attach them to geographic features in a database (Figure 15-3). The old adage "a picture is worth a thousand words" may be modified to "a picture saves a thousand hours." These systems will greatly aid planning, management, and analysis by more easily providing images in GIS. For example, the type, relative location, and condition of public utilities such as fire hydrants may be described with both photos and alphanumeric data collected in a database. If a work order is required to repair a hydrant, a photograph may be taken in the field and tagged to the work order. This photograph may be

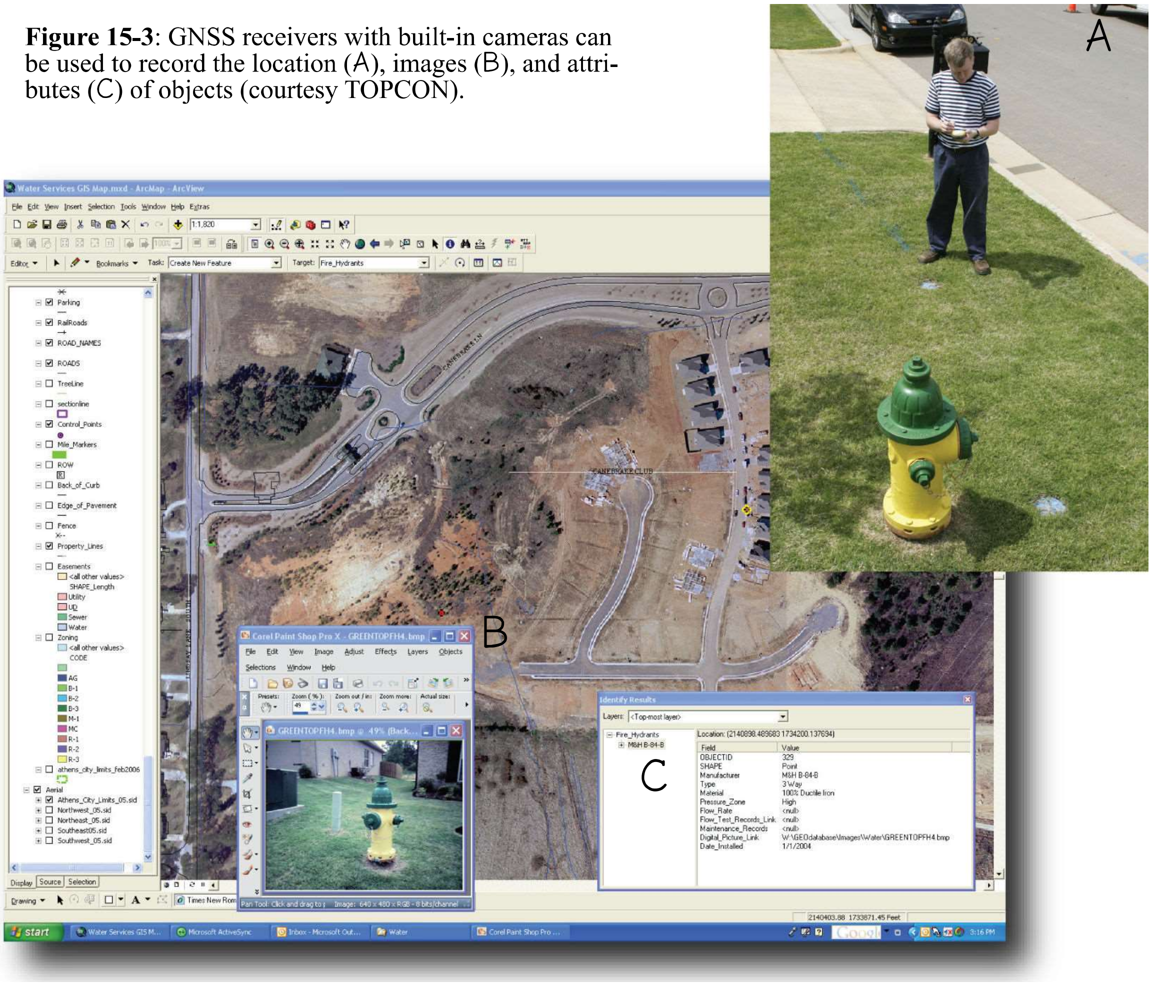


Figure 15-3: GNSS receivers with built-in cameras can be used to record the location (A), images (B), and attributes (C) of objects (courtesy TOPCON).

inspected to verify the type of hydrant, to perhaps identify the tools needed for repair, or to recognize which specific parts are required for maintenance.

Fixed and Mobile Three-Dimensional Mapping

GNSS is also being combined with new advances in ground-based laser scanning to increase the scope, accuracy, and efficiency of spatial data collection. Three-dimensional scanning devices have been developed that measure the horizontal and vertical location of features (Figure 15-4). This scanning is necessary because many features are modified over time, for example, roads are changed, buildings are extended, extra supports may be added to towers, or oil refineries may be re-plumbed. Inventories must be updated to record the features as built, rather than as designed or observed during the previous inventory. A three-dimensional scanning laser may be combined with a precise GNSS receiver to measure the X, Y, and Z coordinates of important features. The GNSS is used to determine the location of the scanning laser. The horizontal and vertical offsets from the scan point are measured by the laser. These measurements are combined with coordinate geometry to calculate the precise positions for all features scanned in the field.

The trend of multi-technology integration will accelerate for rapid, centimeter-level mobile mapping. Multichannel GNSS combined with other positioning systems will provide highly accurate locations, and three-dimensional laser scanners and 360 degree image data collection will allow faster collection of X, Y, and Z coordinates (Figure 15-5).

Combined with GNSS systems and mounted on mobile platforms, three-dimensional laser mapping systems will collect

Figure 15-4: Portable, 3-dimensional scanners are in large-scale production, and allow rapid, accurate collection of x, y, and z positions along with image and other data (courtesy Riegl Systems).

Figure 15-5: Mapping systems combined with GNSS, three-dimensional imaging lasers, optical imaging systems, and other measurement systems to provide integrated three-dimensional measurements in real time (courtesy Apple).

highly accurate data accessible by anyone with a traffic-enabled GNSS. Approaching drivers may be forewarned, travel times calculated and new suggested routes identified. One can imagine self-driving automobiles that navigate via a combined GNSS/LiDAR/GIS, using systems to avoid collisions via real-time distance measurements and wireless communications with "nearby" automobiles.

Such mobile systems will help improve the currency and accuracy of digitized transportation networks, and anything visible from them. Every road can be digitized while driving, as well as every building, light pole, sign, bench, tree, or any other three-dimensional structure visible from them. Efforts will move from a focus on the development of integrated, turn-key data collection systems to software and methods that automate workflow, so that data may travel from the device to the database with as little human intervention as possible.

Automobile systems would rely on multiple technologies. A GNSS would locate the vehicle to within a few tens of centimeters in real time. Three-dimensional data on road centerline, edges, curbs, adjacent poles, and other important features would be identified for the trajectory ahead. When combined with an on-vehicle laser scanner, the system may identify moving automobiles and distinguish them from unexpected stationary objects within the roadway, or other changes in conditions. A combination of LiDAR, RADAR, and cameras may help identify objects, and compare their location to expected features in on-board spatial data. Automobiles may be reliably identified. Aids, such as virtual illumination on the windshield, may be used to highlight other vehicles or road edge when visibility is poor

(Figure 15-6), flag upcoming hazards or turns, or warn of unexpected conditions. The autos could use mapped information on shoulder width, nearby off-ramps, the road ahead, or other structural information to execute the appropriate driving maneuvers and avoid accidents.

Ground Based Positioning

Ground-based positioning systems may find increased adoption. Although GNSS systems provide remarkable positioning information, they are limited. GNSS signals can't pass directly through most solid objects. Buildings, mountains, and dense forest canopies entirely or partially block GNSS signals, yielding a reduced set of observations and reduced spatial accuracy, or at worst, loss of position. GNSS doesn't work in many indoor locations.

Signal strength is one limitation of GNSS systems. Satellite launch constraints force the use of relatively low-power transmitters, and transmission distances are quite large, further dissipating energy. Signals at the receiver are often weak, and difficult to distinguish from multi-path transmissions.

Ground-based positioning services are under development that solve many of these problems. These rely on a set of distributed transmitters and precisely surveyed locations, and are similar to GNSS in using the same basic principle of range measurements and triangulation (Figure 15-7). Each ground-based station transmits a coded signal, which is decoded in a receiver to calculate a range, and then precise location. Centimeter level positioning is possible, and these ground-based systems may be used independently or in conjunction with GNSS positioning. Ground-based antennas transmit signals that are orders of magnitude stronger than GNSS (Figure 15-8). This greatly enhances reception in sub-canopy environments. Since transmitters may be small, dense deployments across high buildings may effectively remove the urban canyon limitations.

Separate systems have been proposed for indoor positioning. These fuse a number of technologies, including Bluetooth and

Figure 15-6: Heads-up displays may appear in autos, improving driver safety and trip efficiency under inclement weather (courtesy NVIDIA).

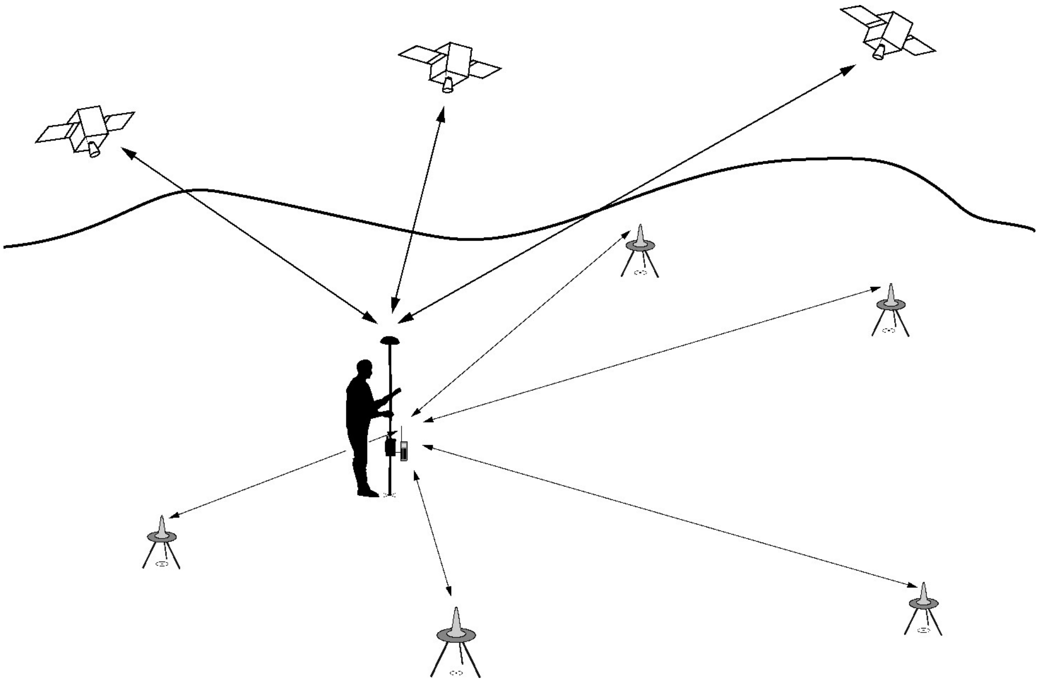


Figure 15-7: A diagram illustrating ground-based positioning systems, here used in conjunction with GNSS satellites. Simultaneous measurements to terrestrial transmitters allow precise positioning, with stronger signals, potentially more robust reception, and the ability to reduce signal obstruction in urban or mountainous environments.

Figure 15-8: A ground-based transmitter and antenna for precise positioning. These GNSS-like systems offer substantial advantages over and complement GNSS systems in some applications (courtesy Locato).

WiFi beacon measurements, and "dead reckoning," using a gyroscope to measure distance and direction traveled from the last known location, and coordinate geometry as described in Chapter 5 to update current position. Married with 3D interior GIS data, they may support many convenience, energy, efficiency, and safety of life applications.

These systems are expected to be more costly on a per-unit area coverage basis than satellite GNSS signals, which are essentially free to the end user. However, ground-based systems may still find application, particularly if robust, centimeter-level positioning is required for self-driving vehicles or autonomous robotic navigation on streets or through buildings.

Datum Modernization

New datum realizations will be calculated for U.S. territories, based on improved measurements and a change in the basic model and methods for datums. Datums in the U.S. have been based on a non-Earth centered ellipsoid to maintain compatibility with previous systems and measurements. The disadvantages of this system now outweigh the advantages, and so a new official datum system is proposed for North America in 2022.

As described in Chapter 3, the continents are moving about the Earth on plates, sometimes at rates exceeding 2.5 cm (an inch) a year. Over several decades, this drift leads to changes in the relative positions among points on different plates (Figure 15-9). In addition, geodesists must factor the total amount of movement into their development of datums, because measurements have been made over several decades, so the relative positions of monuments depend on both the time and location of the respective measurements. Further, the calculations require we establish a stationary reference frame against which to measure points. Because of differences in our starting point, and in how we account for crustal movement, there are large differences between the NAD83 family of datums used primarily in North America, and the ITRF datums used by most of the rest of the World.

The ITRF is an Earth-centered system based on measurements of the X, Y, and Z locations and velocities of points. It places the origin of the adopted ellipsoid at the best estimate of the center of mass of the Earth at the time of each adjustment. The post-1986 NAD83 datums are similar in that they are centered on an Earth model. In contrast to the ITRF system, the NAD83 datums have not adopted the best measurements of the Earth's center, but rather a position compatible with older NAD83 datums. This center assigned a value relative to average crustal velocities on the North American tectonic

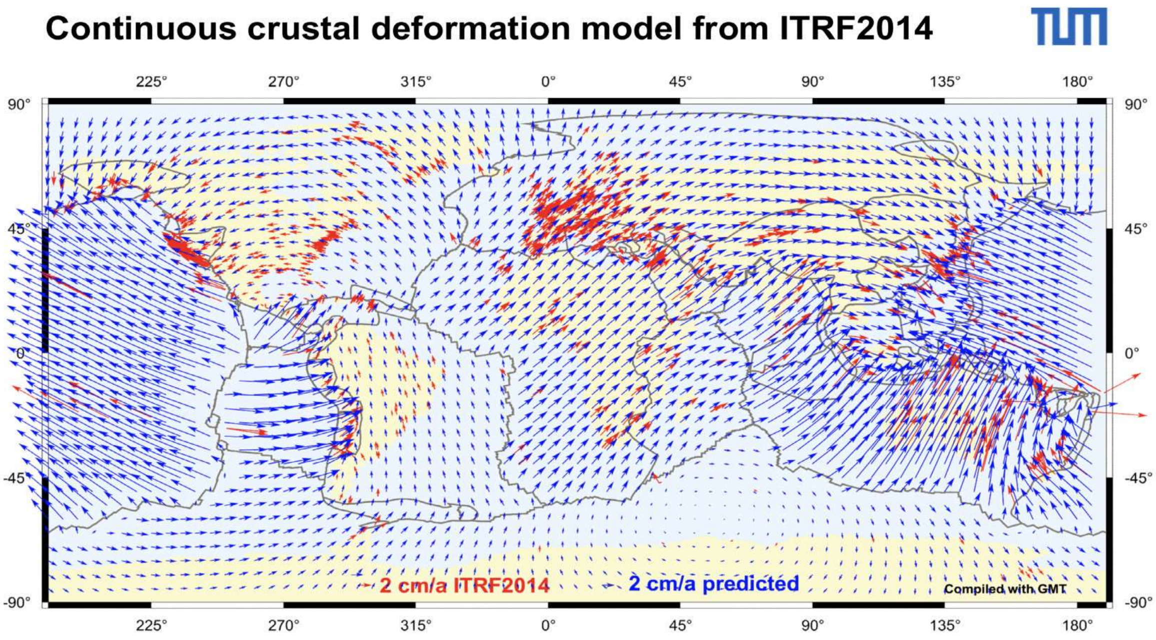


Figure 15-9: Velocity vectors of the Earth's surface, measured at ITRF stations across the globe. Note the relatively high velocities of the eastern Pacific Plate and various different directions of overall plate travel. These movements must be factored into future datum realizations (with permission H. Drewes, www.sirgas.org/fileadmin/docs/Boletines/2017_Drewes_ACK_IM_based_on_ITRF2014_IAG_Kobe.pdf).

plate, rather than a global network. There were many good reasons for maintaining the old origin, primarily because it maintained coordinate compatibility with older NAD83 datums, could be deployed rapidly, and was integrated with a relatively dense network of continuously operating stations within North America.

With the development of the ITRF, the active participation of the NGS, and the integration of CORS stations into the ITRF network, there is strong impetus to harmonize datums in North America with international efforts. Improved global measurements will support more accurate horizontal and vertical datum development, and help support precise, rapid, centimeter-level positioning worldwide.

As noted in Chapter 3, there is a plan to update the datums used in North America, with the introduction of the *North American Terrestrial Reference Frame of 2022* (NATRF2022). This will initally remove much of the positional difference between ITRF/WGS84 and U.S. datums. It will entail a shift, often up to two meters (six feet) in NAD83(2011) coordinates to NATRF2022 coordinates.

Current plans fix the NATRF2022 to the North American and related tectonic plates. This will make position shifts small due to continental drift, and so generally result in stable locations across time over most of North America. This is different than the ITRF system, which generally holds the average shift across all tectonic plates to be zero, thereby distributing shifts across the globe. The ITRF and NATRF2022 positions will diverge over time. The datum transformation between these two systems will be produced, and time dependent. It will likely be in the form of an initial X, Y, and Z transformation, and then a time-dependent piece in which the time difference between input and output datums is accounted for, to incorporate relative continental drift.

The ten-year plan also describes the process for improving the vertical datum for North America. Improvements will be based on gravity measurements across the hemisphere, accounting for changes in gravity fields over time, and also for the rises in mean sea level. Much as with the horizontal datum, it will be based at least in part on an integrated, global set of satellite-based measurements, and tied to ITRF measurements. The new vertical datum will allow the calculation of heights tied to a measurement epoch.

Improved Remote Sensing

Spatial data collection will be substantially improved with the continuing advances in remote sensing. More satellites, higher spatial and temporal resolution, improved digital cameras, and new sensor platforms will all increase the array of available data. We will be able to sense new phenomena, and locate previously measured features with increased precision and accuracy. Satellite-based systems will continue to increase in resolution and coverage, in particular the frequency of data collection. The Worldview system is a salient example of this trend (Figure 15-10). Three satellites have been launched, the latest with a 0.25 m spatial resolution. This is better than most midscale aerial photographs of a decade ago. Similar improvements in resolution and coverage are in progress for other satellite image providers, increasing the frequency and types of images available for medium- to high-resolution mapping.

Parallel improvements continue in aerial image acquisition. Aerial cameras increase in spatial resolution, meaning increasing availability of detailed images, with higher radiometric sensitivity, yielding a broader range of applications. Many systems have higher radiometric breadth, leading to routine collection of more than the visible wavelength spectrums. National aerial acquisition programs are integrating these improvements, with NASS images commonly provided at a one-meter resolution

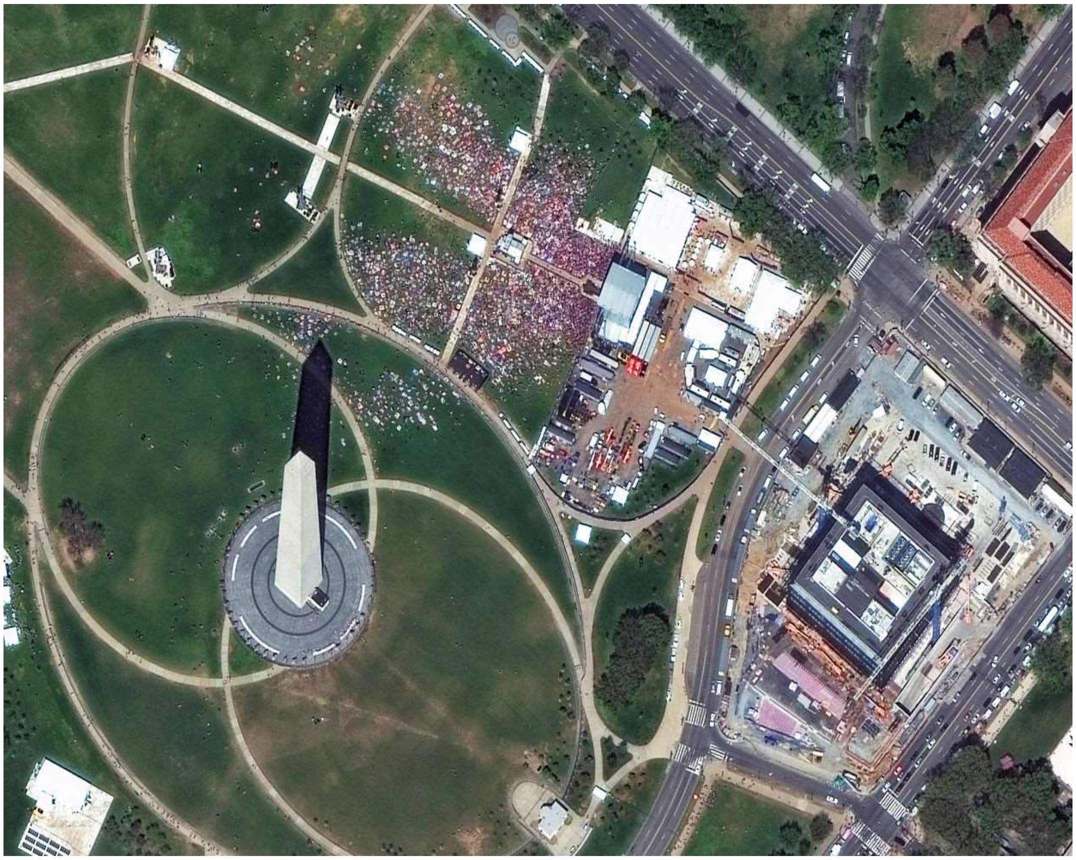


Figure 15-10: An image from the Worldview-3 satellite, taken over Washington, D.C. This satellite produces up to 25 cm resolution images, and is soon to be joined by others of the same high resolution. We may soon have daily, global, high-resolution images (courtesy DigitalGlobe).

image, up from the common two-meter resolution a few years past. The USGS and other organizations are providing subfoot resolution images, perhaps nationwide. Individual light poles, curbs, and even parking lot cracks may be observed in these images, rendering them a rich source of spatial data.

LiDAR data are another example of improved remote sensing. LiDAR systems are increasing in accuracy and resolution and declining in cost, with coordinate data commonly paired with digital aerial images. Commercial LiDAR systems in the recent past collected a data point every few square meters, while current systems routinely collects several samples per square meter. Soon, tens or hundreds of samples per square meter will be common, allowing unprecedented spatial definition.

LiDAR data are dropping in cost, and county to statewide LiDAR mapping will likely become common. Fusion of LiDAR data with other image and spatial data will continue, and surely create new opportunities and applications.

Three-dimensional GIS will be more widely developed and practiced due to data provided by three-dimensional, ground or near-ground level LiDAR (Figure 15-11). LiDAR systems carried on foot, on autos or drones, or in high-flying aircraft will provide feature X, Y, and Z coordinates from various vertical, oblique, and horizontal perspectives. When combined these allow truer three-dimensional characterizations of space. Data development, management, analysis, and visualization is currently taking place across architecture, CAD, surveying, and GIS softwares and disciplines, and substantial development and fusing across these disciplines is in the offing.

Advances in full-sized and miniature aircraft (Figure 15-12) are leading to increased availability of a broader range of aerial imagery. Pioneered primarily by NASA decades ago, this technology is making the leap to broad commercial application. Some experimental UAVs may fly faster and turn tighter than many planes carrying human pilots. Specialized payloads may be carried cheaply on these crafts, for

Figure 15-11: A photograph (upper left) and three-dimensional LiDAR data set, recording precise coordinates at a very high resolution. These coordinate data are useful for inspection, engineering, and maintenance, and were usually unavailable, or quite expensive, prior to LiDAR (courtesy Landpoint).

Figure 15-12: UAVs fitted with cameras and advanced GNSS may collect sub-centimeter accuracy spatial data from LiDAR and aerial imagery (courtesy Geodetics and Velodyne).

long periods of time, and in more dangerous conditions than in human-piloted aircraft. Small craft may be deployed quickly, and have demonstrated ultra-precise three-dimensional measurements at very high resolutions and for site or small, area-specific applications.

There is currently intense development of small helicopters or airplanes outfitted with cameras, LiDAR, and GNSS position, control, and telemetry electronics. This allows preprogrammed flight paths and on-board "intelligence" to modify steering in response to wind, rain, or other in flight conditions. UAVs have completed transoceanic flights with guidance only at takeoff and landing, and UAVs are routinely collecting weather, water quality, and other environmental data.

There have been parallel efforts to reduce the size, weight, and improve the accuracy of LiDAR systems, as most current sensor systems are too large and heavy for most small UAVs. LiDAR system weight increases UAV size, limits collection time, and increases the hazards of collection, but there has been great progress in system miniaturization, which should continue. UAVs will increasingly be used to provide images for coordinate and attributed data collection,

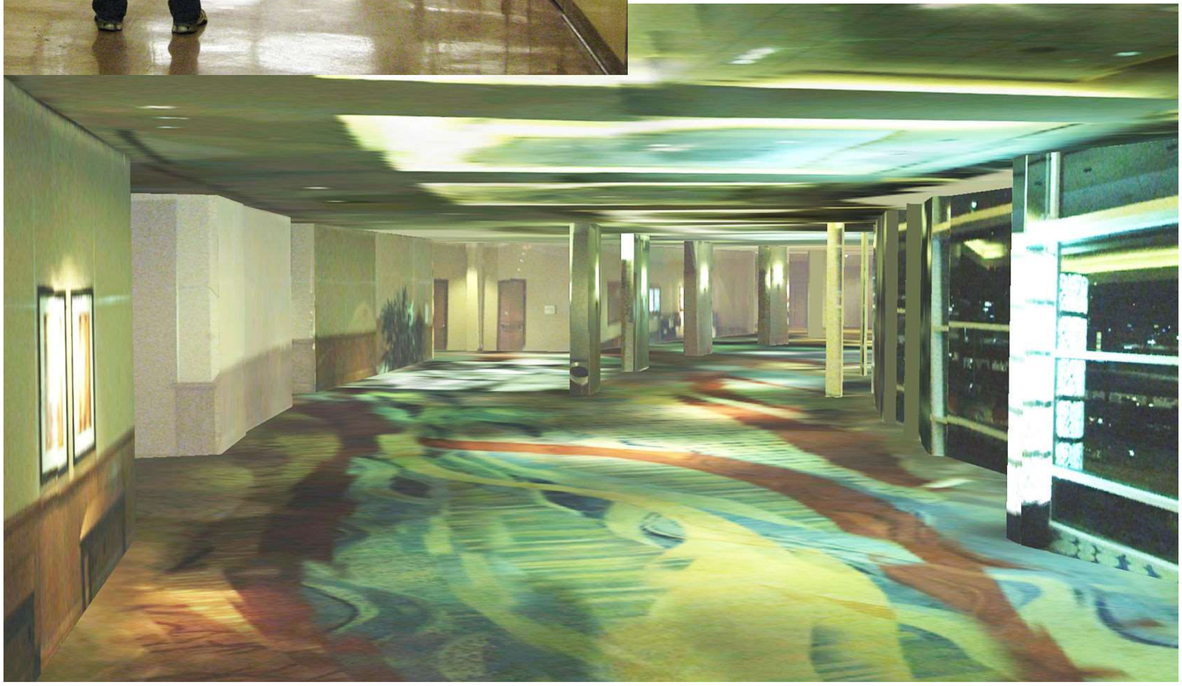
including three-dimensional data, and images from the sides as well as tops of buildings or other structures (Figure 15-13).

Remote sensing is also venturing indoors, enabling data collection of building interiors. These data are useful for maintenance, security, planning, emergency evacuation, and damage assessment.

Experimental, backpack mounted data acquisition systems collect positions via LiDAR and color and texture from multiple cameras (Figure 15-14). These data may be stitched together to provide full three-dimensional representations, and combined with exteriors to create a full coordinate and color record of building interiors and exteriors.

Software vendors are developing data models and workflows to use these data. For example, ESRI has introduced an indoor mapping product that supports development, analysis, planning, and output in a 3D building model. These provide more than virtual walk-throughs, as useful as these are, to allow distance, area, and volume measurement, fixed and movable asset management, and engineering, environmental conditioning, and remodeling calculations.

Figure 15-13: An example of extremely high-resolution LiDAR and image coordinate data. This is a combination of multiple image views to create a three-dimensional data set of an area on the U.C. Berkeley campus. These are real data, and not generated animations, created from subcentimeter-resolution X, Y, Z data. Image data provide color information and LiDAR data location an texture for every point, allowing lifelike renditions from any location. Research seeks to overcome data collection, storage, and processing limitations, so data such as these are routinely used in the not-too-distant future (courtesy Avideh Zakhor, U.C. Berkeley Video and Image Processing Lab).



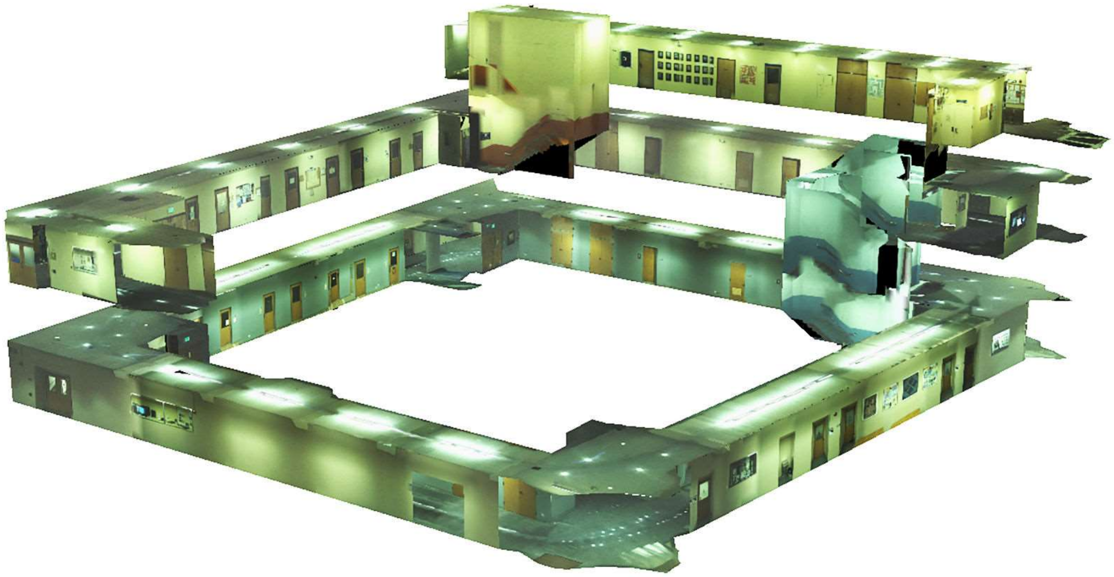


Figure 15-14: An example of indoor, three-dimensional data acquisition and organization. Image and LiDAR data collected with a back-pack system (top) provide three-dimensional reconstructions of indoor environments (middle). These may be combined, maintaining accurate absolute position, to reconstruct connected building interiors (bottom). While currently an active research area, progress is apparent in applications, and will accelerate as data collection, storage, processing, and analysis hurdles are overcome, and as standards are established for data storage and access (courtesy Avideh Zakhor, U.C. Berkeley Video and Image Processing Lab).

Cloud-Based GIS

Cloud-based GIS provides data storage, analysis, and display capabilities over the internet, with services usually provided from one or multiple remote locations via the internet. Cloud-based computing includes broadly accessible internet mapping applications, but also includes data storage and a full suite of software-supported analytical capabilities, something that has generally followed a local or private network architecture. During the recent past, GIS software primarily resided on a local or closely networked hard disk, and ran on the central processing unit of a local computer. You downloaded data to a local hard drive, and purchased software to physically install on the local computer, although the software may have referenced licenses or other resources on a (usually) proprietary server on another computer. Cloud-based computing envisions many of these resources provided from distant sources, with the local computers perhaps serving only as a display and command entry portal — data, software, and processing may all be elsewhere.

Cloud computing has many potential advantages. There may be a lower total cost of ownership, because you may only need to use a set of software occasionally, and can pay as you need it, rather than a fixed price irrespective of total use. Economies of scale in data storage and maintenance or in computing power may be favorable, as well as the centralization of specialized technical support. Additional capacity may be added as needed, as market share grows, or specific project demands increase. Resources may be scaled up or back as needed. New capabilities may be rented or tested more easily, and software functions may better adopt a rental model and pricing structure.

Cloud computing may also provide faster, broader, safer, and more continuous data access. Internet connections are increasing in speed, and solid-state memory in large installations provides faster access yet. Large server facilities may be on continuously, always accessible. Large server arrays may be outfitted with proper backup and protection, including data mirroring at distinct locations. Mirroring provides data redundancy, because the same data are stored concurrently at different physical facilities, often miles or even countries apart. If a fire, flood, or other disaster befalls one data server, the mirror image is likely to remain intact.

Internet mapping is perhaps the simplest and most common form of cloud-based computing. Many internet applications allow users to compose maps on a Web page. The individual user has some control over the data layers shown, the extent of the mapped area, and the symbols used to render the map. The internet is different from other technologies because it allows a wide range of people to custom-produce maps. Each user may choose her own data and cartographic elements to display. The user is largely free from any data development chores and thus needs to know very little about data entry, editing, or the particulars of map projections, coordinates, or other details required for the production of accurate spatial data. Typically, the map itself is the end product, and may be used for illustration, or to support analyses that will be performed entirely within the user's head.

These internet mapping applications are particularly appropriate when a large number of users need to access a limited number of data layers to compose maps. The internet users may select the themes, variables, and symbolization, in contrast to a static map graphic, in which a website cartographer defines the properties of each map.

Because most internet mapping applications are built for users who have little knowledge of spatial data, maps, and analysis, the suite of spatial operations allowed is usually very sparse. Most internet mapping is currently limited to creating simple displays. This is changing, as query, distance

functions, and basic tools are provided, albeit in very simple forms.

As noted in Chapter 7, web mapping services are another step toward a cloud-based GIS model. Data are stored somewhere "on the cloud," and a specific link to them provided. Data are accessible after forging a connection, as if they were from any other disk source, at least as far as the accessing software view. The GIS program doesn't distinguish between local and cloud data once the connection is made.

This brings up one major limitation of the cloud model, its dependence on a fast connection to the cloud of resources. Slow internet speeds become a hindrance, particularly with large data or image sets that characterize many spatial analyses. Each zoom, pan, or layer addition may require a scene to be repainted, involving the movement of billions of pixels through the web connection. While this may be overcome to some extent with local caching, anticipatory downloading, for example, pulling data in a wider area than that immediately viewed, and other software techniques, but within limits. In many instances there is no substitute for extremely fast internet speeds. Widespread, fast internet should be forthcoming, but access may depend on internet demand as well as supply.

To date, analytical tools delivered over the internet are still quite rudimentary, and may likely remain so for some time. Robust, correct operation of an analytical tool is difficult to provide in many instances, and requires a sizeable investment. Systems for delivery, payment, and protection for a broad, interacting suite of geospatial tools will take development, both technically and culturally.

Open GIS

Open Standards for GIS

Open standards in computing seek to reduce barriers to sharing programs, data, and information. Spatial data structures may be very complex, perhaps more than many other kinds of data. Data may be raster or vector, real or binary, or represent point, line, or area features. In addition, different software vendors may elect to store their raster imagery using different formats, and data may be delivered on different physical media, or formatted different ways. If a person orders an image in one format, but her computing system does not support the physical media on which the data are written, or does not understand the file structures used to store the image, then she may not be able to use these data. Incompatible systems are generally described as non-interoperable, and open standards seek to remove this non-interoperability.

The development of open standards in computing is driven by the notion that the larger user community benefits when there are no technical barriers that inhibit the free exchange of data and methods. Open standards seek to establish a common framework for representing, manipulating, and sharing data. Open standards also seek to provide methods for vendors and users to certify compliance with the standard. Standards have been developed in a number of endeavors; for example, the ISO 9600 specifications for physical storage formats allow any manufacturer, data developer, or user to build, read, write, or share data on hard drives, optical disks, tapes, or other storage devices.

Businesses and many other organizations by their nature have a proprietary interest in the spatial data entry, storage, and methods they produce. Many vendors survive by the revenue their GIS products generate, and so have a strong interest in protecting their investments and intellectual property. However, the developers also may

spur adoption of their GIS packages and speed up the development of complementary software by making the internal workings of some portions of their GIS packages public knowledge, for example, by publishing the data structures and formats used to store their spatial data. Thus, these vendors also have a strong interest in making parts of their system open to the public.

Open standards for spatial data are the responsibility of the OpenGIS Consortium. The OpenGIS Consortium has developed a framework to ensure interoperability. They do this by defining a general, common set of base data models, types, domains, and structures, a set of services needed to share spatial data, and specifications to ease translation among different representations that are compatible with the OpenGIS standards. Data developed by a civil engineer and stored in a raster format on a Unix version of Arc/Info should be readily accessible to a soil scientist using GRASS on an OS-X Apple system.

Open standards in GIS are relatively new. While most of the large software vendors, data developers, and government and educational organizations are members of the OpenGIS Consortium, some components of the standard are still under development. In the future, there will be increased emphasis on compliance to the OpenGIS standards.

Open Source GIS

Open source software is different from most other software in that it is distributed free, along with the source code. The open source organization (www.opensource.org) requires that the software is not by design restricted to a specific operating system or other technology, that there can be no royalties, and that there be no explicit discrimination against fields of endeavor, persons, or groups. But the main, defining characteristic of open source software is an open, grassroots network of collaborators developing, documenting, and freely sharing source code.

There are open source software of many types, from operating systems to word processors, and including GIS. Open source GIS software projects are directed at a range of applications, and notable examples include the development of general-purpose GIS (e.g., GRASS, FMaps) to specific utilities (e.g., MapServer for Web-based spatial data display, query, and analysis) or toolkits to support GIS software development (e.g., GDAL, shapelib).

Open source use is a large and growing phenomenon for many reasons. High software costs are driving many organizations toward open source software. Licenses for some commercial products are tens to hundreds of thousands of dollars annually for some large organizations. If these organizations employ staff programmers, open source GIS may meet geoprocessing needs at a reduced cost.

Many organizations use open source GIS because commercial products may not provide the required functions or capabilities. Three-dimensional structural analysis tools may exist that meet the requirements of a mining company, and so they may develop specific applications. This development may be more efficient and less expensive in an open source environment.

Open source use is expanding in many countries because of specific governmental initiatives. China, India, and Brazil have all supported open source software in general, and operating systems in particular, to maintain independence from foreign firms, reduce costs to government and local business, and develop local information technology expertise. Because these nations are home to more than a third of the world's population, their actions alone are substantially increasing the use of open source GIS.

A Hybrid Model

Proprietary software vendors may adopt a hybrid software approach, where they interact with open software and systems. This has taken many guises. Some may simply support standards, and ensure their sys-

tems may access and generate industry standard data forms. But a fuller approach provides the code in a mix of open and proprietary parts. Base code may be provided free, with a charge for extensions or some set of additional capabilities.

Alternately, there may be charges for the base code, but enough source code or adherence to open standards that open source extensions can be easily added later. This allows for the development of an "ecosystem" of extension around a base application, both proprietary and open source.

Summary

GIS are a dynamic collection of conceptual models, tools, and methods that use spatial data. As such, they will continue to evolve. What becomes standard practice in the future may be quite different from the methods we apply today. However, the fundamental set of knowledge will remain unchanged. We will still gather spatial and attribute data, adopt a spatial data model to conceptualize real world entities, and use map coordinates to define positions in space. The coordinates are likely to remain based on a standard set of map projections, and we will combine the spatial data of various classes of entities to solve spatial problems. This book is an attempt to provide a foundation to effectively use spatial analysis tools. I hope it has provided enough information to get you started, and has sparked your interest in learning more.

Suggested Reading

The World Wide Web is the best source for information about new developments and trends in spatial data acquisition, analysis, and output. In contrast to previous chapters, nearly all the suggested readings are websites. We apologize that many links may be short-lived, but the reader is directed to search for similar and additional sites for the most current information.

www. gislounge.com

www.nasa.gov, general NASA entry point

www.gis.com, an ESRI-sponsored website, general information

www.usgs.gov, public domain data from the USGS

www.usgs.gov/centers/eros, another common USGS entry point for image and GIS data

www.epa.gov/geospatial/

www.opengeospatial.org, open GIS consortium

https://catalog.data.gov/dataset/lidar-point-cloud-usgs-national-map, LiDAR data and systems description

www.gpsworld.com/

www.digitalglobe.com, high-resolution satellite data

www.gisuser.com

www.directionsmag.com

Study Questions

15.1 - Which of the described new technologies is likely to have the largest impact in GIS over the next five years? Why?

15.2 - What are areas of spatial data entry, analysis, output, or storage that are in dire need of innovation or new and better methods? What is a major bottleneck to further advancement of spatial information science and technology?

15.3 - What is Open Source GIS? How will this change spatial computing?

Appendix A: Glossary

Terms used in GIS and Spatial Data Development and Analysis

Accuracy: The nearness of an observation or estimate to the true value.

Active remote sensing system: A system that both emits energy and records the energy returned by target objects.

Adaptive sampling: A method to increase sampling efficiency by increasing the spatial sample frequency in areas with higher spatial variability.

Adjacency: Two area objects that share a bounding line are topologically adjacent.

Affine coordinate transformation: A set of linear equations used to transform from one Cartesian coordinate system to another. The transformation applies a scaling, translation, and rotation.

Almanac: Important system information sent by each GPS satellite, and recorded by a GPS receiver to obtain current satellite health, constellation status, and other information helpful for GPS positioning.

Application server: a middle tier in a common database architecture, that passes requests for data from higher tier, user interfaces, to the database storage and server tier at lower levels.

Arc: A line, usually defined by a sequence of coordinate points.

ArcGIS: A GIS software package produced by Environmental Research Systems, Inc., of Redlands, California.

Area feature: A polygon, collection of contiguous raster cells, or other representation of a bounded area. The feature is characterized by a set of attributes and has an inside and an outside.

ASCII: American Standard Code for Information Interchange. A set of numbers associated with a symbol used in information storage and processing. Numbers are between 0 and 255 and may be represented by a single byte of data.

Aspect: The direction of steepest descent on a terrain surface.

Atmospheric delay: A change in the speed of light, and more specifically GNSS signal speed, when passing through the atmosphere.

Atmospheric distortion: Image displacement due to the bending of light as it passes through the atmosphere.

Attribute: Non-spatial data associated with a spatial feature. Crop type, value, address, or other information describing the characteristics of a spatial feature are recorded by the attributes.

Autocad Geospatial: A suite of GIS software systems produced by Autodesk, Inc., of San Rafael, California.

AVHRR: Advanced Very High Resoluton Radiometer. A discontinued satellite system run by the National Oceanographic and Atmospheric Administration to collect visible, thermal, and infrared satellite images of the globe each day. The system had up to a 1.1 km resolution, and was the earliest satellite with daily global coverage, useful for landuse mapping.

Bandwidth: A parameter used in kernel mapping to affect the influence, or spread, of each point observation.

Base station: GNSS recording station over a precisely surveyed location, used in differential correction.

Beacon receiver: A GNSS receiver capable of decoding beacon base station signals transmitted by the U.S. Coast Guard beacon stations.

Bearing: A direction, usual specified as a geographic angle measured from some base line, e.g., true north.

Beidou: A Chines GNSS systems

Benchmark: A monumented, precisely surveyed location for which coordinates are known to a high degree of accuracy.

Bilinear interpolation: A method for calculating values for a grid location based on a linear combination of nearby grid values.

Binary classification: A classification of spatial objects into two classes, typically denoted by a 0 class and a 1 class.

Binary operation: A spatial operation with two inputs.

Bit: A binary digit. A bit has one of two values, on or off, zero or one. This is the smallest unit of digital information storage and the basic building block from which all other computer data and programs are represented.

Boolean algebra: Conditions used to select features with set algebraic conditions, including and, or, and not conditions.

Boundary generalization: Simplification of the “true” boundary lines that define features due to the inability to record every point along a boundary. Some sampling must occur in any non-mathematically defined boundary, and a straight line segment often does not fully represent curves between the endpoints.

Buffer: A buffer area is a polygon or collection of cells that are within specified proximities of a set of features. A buffer operation is one that creates buffer areas.

Bundle adjustment: The simultaneous removal of geometric distortion and production of orthophotographs from a number of aerial images.

Byte: A unit of computer storage consisting of 8 binary digits. Each binary digit may hold a zero or a one. A byte may store up to 256 different values.

C/A code GPS: Coarse acquisition code, a GPS signal used for rapid, relatively low-accuracy positional estimates. Accuracies without further corrections are typically from a few to tens of meters.

CAD/CAM: Computer Aided Design/Computer Aided Mapping. Software used primarily by design engineers and utilities managers to produce two and three dimensional drawings. Related to GIS in that coordinate information is input, manipulated, and output. These systems often do not store map-projected coordinates, and do not have sophisticated attribute entry and manipulation capabilities.

Cadastral: With reference to property lines or ownership, for example, a cadastral layer usually contains property lines, and a cadastral survey is the survey of property lines.

Candidate key: A column or columns in a relational table that meets the requirements for a key, primarily that it uniquely identifies every row in the table.

Carrier-phase GPS: Relatively slow but accurate signal used to estimate position. Position may be determined to within a few centimeters or better.

Cartesian coordinate system: A right-angle two or three-dimensional coordinate system. Axes intersect at 90 degrees, and the interval along each axis is linear.

Cartographic modeling: The combination of spatial data layers through the application of spatial operations.

Cartographic object: A digital representation of a real-world entity.

Cartometric map: A map produced such that the relative positions of objects depicted are spatially accurate, within the limits of the technology and the map projection used.

Cell dimension: The edge length of square cells used in raster data sets.

Centroid: A central point location for an area feature, often defined as the point with the lowest average distance to all points that define the area boundary.

Characteristic hull: A polygon boundary that attempts to include the densest concentration of a set of observed points. Often used in home range analysis, it starts with a Triangulated Irregular Network, or TIN, of points, and winnows large triangles to some specified reduction threshold.

Choropleth map: A map of polygons with colors assigned in a gradient that depicts classified levels of a variable. These are commonly used for population density, average income, health risk, or other variables mapped by administrative boundaries in which high to low categories are of interest.

Classification: A categorization of spatial objects based on their properties.

Clients: Programs that request data from a server.

Clip (overlay): The vertical combination of two data layers, with a clip layer typically designated that defines the extent and location of output areas, and that preserves only the data from the non-clip layer for the clip area.

Cloud computing: Utilizing computer processing that is remotely located, typically on a networked array of distant computers, accessed through the World Wide Web.

Cluster sampling: A technique of grouping samples, to reduce travel time among samples while maintaining sample number.

Code-phase GPS: see C/A code.

COGO: Coordinate Geometry, the entry of spatial data via coordinate pairs, usually obtained from field surveying instruments.

COMPASS: Chinese satellite-based positioning systems.

Concatenated key: The use of two or more table columns as a key in a relational database management system.

Conformal coordinate transformation: A registration that requires scale changes to be equal in the x and y directions.

Conformal projection: A map projection is conformal when it preserves shape for some portions of the map.

Conic projection: A map projection that uses a cone as the developable surface.

Connectivity: A record or representation of the connectedness of linear features. Two linear features or networks are connected if they may be traversed without leaving the network.

Continuous surface: A variable or phenomenon that changes gradually through two-dimensional space, e.g., elevation or temperature.

Contour line: A line of constant value for a mapped variable.

Control points: Point locations for which map projection and database coordinate pairs are known to a high degree of accuracy. Control points are most often used to convert digitized coordinates to standard map projection coordinates.

Convergent circle: A circle used in defining a facet for a triangulated irregular network, that passes through three points, and does not contain any other points.

Convex hull: The polygon that completely contains a set of points and that has no acute (< 180°) exterior angles.

Coordinates: A pair or triplet of numbers used to define a position in space.

Coordinate transformation: The conversion or assignment of coordinates from a non-projected coordinate system to a coordinate system, typically via a system of linear mathematical equations.

Core area: The central or primary concentration for a set of points.

Cost surface: A spatial depiction of the cost of traveling among locations in an area.

Cubic convolution: A method of calculating grid values based on a weighted combination of 16 nearby grid cells.

Cylindrical projection: A map projection that uses a cylinder as the developable surface.

Dangle: An unintended overshoot in a line segment when crossing another line segment.

Data independence: The ability to make changes in data structure in a database management system that are transparent to users or applications that use data.

Data model: A method of representing spatial and aspatial components of real-world entities on a computer.

Database management system (DBMS): A collection of software tools for the entry, organization, storage, and output of data.

Datum: A set of coordinate locations specifying horizontal positions (for a horizontal datum) or vertical positions (for a vertical datum) on the Earth surface.

Datum adjustment: A re-calculation of a datum based on additional measurements.

Datum realization: The outcome of a datum re-adjustment, a specific, defined datum surface and set of datum points.

Datum shift: The change in horizontal or vertical point location that results from a datum adjustment.

Datum transformation: A method or set of equations that allows the calculation of a point location in a one datum based on coordinates expressed in a different datum.

Declination: The angle between the bearing towards True North and the bearing towards Magnetic North.

Delaunay triangles: The set of triangles formed in a triangulated irregular network, connecting points to the nearest points to create triangles, while ensuring that the triangle edges don't cross, and are formed by convergent circles.

DEM: Digital Elevation Model, a raster set of elevations, usually spaced in a uniform horizontal grid.

Developable surface: A geometric shape onto which the Earth sphere is cast during a map projection. The developable surface is typically a cone, cylinder, plane, or other surface that may be mathematically flattened.

Digital terrain model: A digital representation of elevation, including DEMs, TINs, and other digital representations.

Diaphragm: A camera component that functions like the iris of the human eye, to control the amount of light available to fall on the film or CCD recording surface, and to improve focus.

Differential GNSS: GNSS positioning based on two receivers, one at a know location and one at a roving, unknown location. Data from roving receivers are corrected by the difference error computed at the known location.

Digitize: To convert paper or other hardcopy maps to computer-compatible and stored data.

Digitizing table: A device with a flat surface and input pointer used to digitize hardcopy maps.

Dilution of precision (DOP): See position dilution of precision (PDOP).

Dissolve: An operation that removes lines separating adjacent polygons that are considered equal, based on some characteristic or measure. A dissolve operation is typically applied based on equal values of variables that are contained in a table associated with the data layer.

DLG: Digital Line Graph, vector data developed and distributed by the United States Geological Survey.

Domain: The range of values a variable may take.

DOQ: Digital Orthophoto Quadrangle, an orthographic photograph provided in digital formats by the USGS. Most tilt and terrain error have been removed from DOQs.

Dot density maps: Maps with dots placed inside of polygons in proportion to numeric value of a variable. These are used to represent population for counties or states, or other numeric data corresponding to defined areas, with more dots for higher levels of the variable, for example, one dot per 1,000 inhabitants.

DRG: Digital Raster Graphics, a digital version of USGS fine- to medium-scale maps.

Dual-frequency GPS receiver: A receiver capable of measuring the L1 and L2 broadcast signals, and using these to estimate highly accurate and precise positions, typically to centimeter levels.

Easting: The axis approximately parallel to lines of equal latitude in UTM and a number of other standard map projections.

Electromagnetic spectrum: A range of energy wavelengths, from X-rays through radar wavelengths. The electromagnetic spectrum is typically observed at wavelengths emitted by the Sun or by objects on Earth, covering wavelengths from the visible to the thermal infrared region.

Ellipsoid: A mathematical model of the shape of the Earth that is approximately the shape of a flattened sphere, formed by rotating an ellipse.

Ellipsoidal height: Height measured from an ellipsoidal surface to a point on the surface of the Earth.

Endlap: The end-to-end overlap in aerial photographs taken in the same flight line.

Entity: A real world item or phenomenon that is represented in a GIS system or database.

Ephemeris: Information on GNSS satellite orbits, required by GNSS receivers to computer satellite position, range distance, and receiver position.

Epsilon band: A band surrounding a linear feature that describes the positional error relative to the feature location.

Equal-area classification: A classification method that assigns classes such that each class corresponds to an equal area.

Equal-interval classification: A classification method that assigns an equally spaced set of classes across the range of a variable.

Erase function: A vector spatial operation, typically that “clips out” and discards the area in one layer corresponding to the polygon boundaries in another data layer.

ERDAS: A GIS and remote sensing image processing software package owned and developed by Leica Geosystems, St. Gallen, Switzerland.

ETM+: Enhanced Thematic Mapper, a scanner carried on board Landsat 7, providing image data with resolutions of 30 meters for visible through mid infrared, 15 meter panchromatic, and 60 meter for thermal wavelengths.

Facet: A triangular face in a TIN.

False northing: A number added to coordinates in a map projection, usually to avoid negative coordinate locations within the area of a map projection.

Feature: An object or phenomenon in the landscape. A digital representation of the feature is often called a cartographic feature.

Feature generalization: The incomplete representation of shape defining coordinates for entities represented in a GIS.

Fiducial marks: Also known as fiducials, precisely scribed marks that are recorded near the edges of aerial images, and used to remove systematic camera distortion and to register images.

FIPS: Federal Information Processing Standards code - a set of numbers for defined political or physical entities in the United States. There are FIPS codes for each state, county, and other features.

First Normal Form: A set of requirements for a relational database table, primarily that there be no repeat column, defined as columns that represent the same kind of information. For example, a database table intended to represent families, repeat columns for children would violate the requirements for first normal form.

Friction surface: A raster surface used in calculating variable travel costs through an area. The friction surface represents the cost per unit distance to travel through a cell.

Flatbed scanner: An electronic device used to record a digital image of a hardcopy map or image.

Flow direction: The direction water will flow from a point, usually an azimuth or bearing angle assigned to a raster cell.

Foreign key: A column in a relational database table that is a candidate key, and used to join the table to a different relational table.

Friction surface: See cost surface.

FTP: File Transfer Protocol, a standard method to transfer files across a computer network.

Functional dependency: Property of a set of items in a database table. If one item is functionally dependent on another, that means knowing the value of one item guarantees we know the corresponding value of the second item.

Galileo: A European-based GNSS system.

Generalization: The simplification of shape or position that inevitably occurs when features are mapped.

Geocentric: A measurement system that uses the center of the Earth as the origin.

Geocoding: The process of assigning a geographic or projection coordinate to a data item that is based on a street address, town, and state or country.

Geodetic datum: A reference system against which horizontal and/or vertical positions are defined. It typically consists of a sphere or ellipsoid and a set of point locations precisely defined with reference to that surface.

Geodesy: The science of measuring the shape of the Earth and locations on or in the Earth.

Geoid: A measurement-based model of the shape of the Earth. The geoid is a gravitational equipotential surface, meaning a standard surface of equal gravitational pull. The geoid is used primarily as a basis for specifying terrain or other heights.

Geoidal height: The distance measured normal to the ellipsoid surface from and ellipsoid to a geoid.

Geographic North: The northern axis of rotation of the Earth. By definition true north lies at 90° latitude.

GeoMedia: A GIS software package produced by Intergraph, Inc., of Huntsville, Alabama.

GIS: A geographic information system. A GIS is a computer-based system to aid in the collection, maintenance, storage, analysis, output, and distribution of spatial data and information.

GLM: Geographic Markup Language, a standard method for documenting and transferring spatial data.

GLONASS: Global Navigation Satellite System. A Russian developed and maintained system for coordinate measurement and positioning.

Global operation: A spatial operation where the output location, area, or extent comes from operations on the entire input area or extent.

GNSS: Global Navigation Satellite System. A constellation of satellites plus a ground control segment that allows precise location on or above the Earth. This includes GPS, GLONASS, and other satellite navigation system.

Gnomonic projection: A map projection with the projection center placed at the center of the spheroid.

GRASS: An open-source GIS software system.

Graticule: Lines of latitude and longitude drawn on a hardcopy map or represented in a digital database.

Gravimeter: An instrument for measuring the strength of the gravitational field.

Great circle distance: The shortest distance between two points on the surface of the Earth. This distance follows a great circle route, defined as the route on the surface defined by a plane that intersects the starting and ending point and the center of the Earth.

Grid North: The direction parallel to the northing axis in a projected, Cartesian coordinate system.

Greenwich meridian: The line of equal longitude passing through the Royal Observatory in Greenwich, England. This line was taken as zero, by convention, for the system of longitude measurements for the world.

GRS80: Geodetic Reference Surface of 1980, an ellipsoid used for map projections in much of North America.

Hardcopy map: A map printed on physical media, usually paper.

Height above ellipsoid, HAE: See ellipsoidal height.

Helmert transformation: A method to transform among horizontal datums.

Hierarchical data model: A method of organizing attribute data that structures values in a tree, typically from general to more specific.

High-pass filter: A raster operation that identifies large or high-frequency differences between cells.

Hydrography: Geographic representation of water features.

Hypsography: Geographic representation of height features.

Ikonos: A high resolution imaging satellite system. Ikonos provides 1-meter panchromatic and 3-meter multispectral image data.

Inner join: A combination of two data tables in a database management system based on a key column. The output table combines rows by matching values in the key column, and saves only rows that have matching key values in both tables.

International Terrestrial Reference Frame (ITRF): A geocentric coordinate reference frame that follows an international standard for specifying Earth coordinates. Defines an origin, ellipsoidal shape, and X, Y, and Z coordinate directions.

Ionospheric delay: The change in travel time of an electromagnetic signal when passing through the ionosphere. Most applicable in GIS to uncertainty in GNSS positioning due to uncertain signal travel times.

Instantaneous field of view (IFOV): The area or angle sensed by an imaging system, or sensing component such as the pixel or lens, of the system.

Interpolation: The estimation of variables at unsampled locations from measurements at sampled locations. Interpolation methods are usually understood to use a formula with all parameters that are pre-determined, meaning that parameter values used in the formula do not depend on the data values.

Intersection (overlay): The vertical combination of two data layers, typically restricted to the extent of one data layer but preserving the data contained in both data layers for that extent.

Interval/ratio scale: A measurement scale that records both order and absolute difference in value for a set of variables.

Isopleth map: A map depicting lines of constant value for a variable, also known as a contour map.

Items: Variables or attributes in a data table, typically viewed as the columns of the table. These are the types of essential characteristics used to described each feature in the geographic data set, e.g., area, depth, and water quality for a lakes data set.

Idrisi: A GIS system developed by the Graduate School of Geography of Clark University, Worcester, Massachusetts.

IDW: Inverse Distance Weighted interpolation, a method of estimating values at unsampled locations based on the value and distance to sampled locations.

Infrared image: An image that records reflectance in the near infrared wavelengths, typically including 0.7 to 1.1 micrometers.

JPEG: An image compression format.

Kernel: An arrangement of cells and values used as a multiplication template in raster analysis.

Kernel mapping: A method of identifying core areas, concentrations, or density of occupation based on “stacking” kernels that represent occurrence frequency at observed locations.

Key: An item or variable in a relational table used to uniquely identify each row in the table.

Kriging: An interpolation method based on geostatistics, the measurement of spatial autocorrelation.

Lambert conformal conic: A common, cone-based map projection.

Landsat: A NASA project spanning more than three decades and seven satellites that proved the capabilities of space-based remote sensing of land resources.

Latitude: Spherical coordinates of Earth location that vary in a north-south direction.

Law of sines: A trigonometric relationship that allows the calculation of unknown triangle edge lengths from known angles and edge lengths.

Leveling surveys: Surveys used to measure the relative height difference between sets of points.

Lidar: Laser detecting and ranging, the use of pulse laser measurements to identify the height, depth, or other properties of features.

Linear referencing: See geocoding.

LIS: A Land Information System, a name originally applied for GIS systems specifically developed for property ownership and boundary records management.

Local operation: A spatial operation where the output location, area, or extent comes from operations on that same extent.

Logical model: A conceptual view of the objects we portray in a GIS.

Longitude: Spherical coordinates of Earth location that vary in an east-west direction.

Manifold: GIS software package produced by CDA International, of Carson City, Nevada.

Map algebra: The combination of spatial data layers using simple to complex spatial operations.

MapInfo: GIS software package produced by MapInfo, Inc., of Troy, New York.

Map projection: A systematic rendering of features from a spheroid or ellipsoid representing the 3-dimensional Earth to a map surface.

Mean center: A measure of the central location of a set of objects or observations, based on the mean x and y coordinates for all observations.

Mean circle: An estimate of a core area via a circle centered on the mean center, with a radius derived from the observed points, for example, the standard deviation.

Meridian: A line of constant longitude.

Magnetic North: The point where the northern lines of magnetic attraction enter the Earth. Magnetic North does not occur at the same point as “True” or Geographic North. In the absence of local interference a compass needle points towards magnetic north. The magnetic north pole is currently located in northern Canada.

Metadata: Data about data, that describes the properties of a spatial data set, including the coordinate system, extent, attribute types and values, origin, lineage, accuracy, and other characteristics needed for effective evaluation and use of data.

Metes and bounds survey: A survey method based on distance and sometimes angle measurements from known or monumented points.

Minimum mapping unit (MMU): The smallest area resolved when interpreting an aerial or satellite image, or when mapping area features from a source data set.

Moving window: A usually rectangular arrangement of cells that shifts in position across a raster data set. At each position an operation is applied using the cell values currently encountered by the moving window.

MSS: Multi-spectral Scanner, an early satellite imaging scanner carried by Landsat satellites.

Modifiable areal unit problem: The dependence of aggregate area statistics on the size and shape of the aggregation units.

MODIS: Moderate Resolution Imaging Sensor. A later generation imaging scanner that is part of NASA's Mission to Planet Earth. Provides high spectral resolution, frequent global coverage, and moderate spatial resolution of from 250 to 1000 meters.

Molodenski transformation: A method to transform among geodetic datums.

Monte Carlo simulation: A method of estimating the variability in a process or model by adding small uncertainties to the input data or parameters over thousands of model runs, and observing how these small differences change the outcome.

Multipart feature: A vector feature, usually of polygons, where multiple, separate geographic entities are grouped and treated as one, and associated with one row in the associated attribute table.

Multispectral: An image, film, or system that records data collected from multiple wavebands.

Multi-tier architecture: A database management system design where there are multiple levels of clients above a server.

Nadir point: The point directly below the aircraft, usually near the center of an aerial image.

NAD27: North American Datum of 1927, the adjustment of long-baseline surveys to establish a network of standardized horizontal positions in the early 20th century.

NAD83: North American Datum of 1983. The successor to NAD27, using approximately an order of magnitude more measurements and improvements in analytical models and computer power. The current network of standard horizontal positions for North America.

NASS-CDL: National Agricultural Statistical Service Crop Data Layer, annually produced raster data sets of crop categories for United States farmland.

NAVD29: North American Vertical Datum of 1929, an adjustment of vertical measurements to establish a network of heights in the early 20th century.

NAVD88: North American Vertical Datum of 1988, the successor vertical datum to NAVD29.

Neatline: A line containing all elements that make up a map.

Neighborhood operation: A spatial operation where the output location, area, or extent comes from operations on an area larger than, and usually adjacent to the input extent.

Network: A connected set of line features, often used to model resource flow or demand through real-world networks such as road or river systems.

Network center: A location on a network the provides or requires resources.

NLCD: National Land Cover Data set, a Landsat Thematic Mapper (TM) based classification of landcover for the United States.

NOAA: National Oceanic and Atmospheric Administration, the U.S. government agency that oversees the development of national datums.

Node: An important point along a line feature, where two lines meet or intersect.

Nominal scale: A measurement scale that indicates the difference between values, but does not reflect rank or absolute differences.

Northing: The axis in the approximately north-south direction in UTM and other standard coordinate systems.

Normal forms: A standard method of structuring relational databases to aid in updates and remove redundancy.

N-tuple: A group of attribute values in a database management system.

NWI: National Wetlands Inventory data compiled by the U.S. Fish and Wildlife Service over most of the United States. These data provide first-pass indications of wetland type and extent.

Object: See cartographic object.

Object-oriented data model: A data model that incorporates encapsulation, inheritance, and other object-oriented programming principles.

Open source software: Computer programs that provide the source code to any user, typically easily accessible through a web portal.

Operation, spatial: The manipulation of coordinate or attribute data.

Optical axis: A ray approximately perpendicular to the film or image plane in a camera and parallel to the center of the lens barrel, that may be thought of as the primary direction of incoming light.

Ordinal scale: A scale that represents the relative order of values but does not record the magnitude of differences between values.

Orthogonal: Intersecting at a 90 degree angle.

Orthographic view: Horizontal placement as would be seen from a vertical viewpoint at infinity. There is no terrain or tilt-perspective distortion in an orthographic view.

Orthographic projection: A map projection with the projection center an infinite distance from the map surface.

Orthometric height: Height measured from the Geoid surface to a point on the surface of the Earth.

Orthophotograph: A vertical photograph with an orthographic view. Orthophotographs are created by using projection geometry and measurements to remove tilt, terrain, and perspective distortion from aerial photographs.

Outer join: A combination of two data tables in a database management system based on a key column. The output table appends those rows in a second table that match values in the key column. Null values are placed in joined-table columns from the second table where there is no match to the first table.

Overlay: The “vertical” combination of two or more spatial data layers.

Overshoot: A digitized line that extends past a connecting line.

Panchromatic: An image, film, or system that records in only one wavelength band, and resulting in gray scale (black and white) images.

Parallax: The relative shift in position of features due to a shift in viewing location.

Passive remote sensing system: A system that does not emit the radiation it records from target objects.

PDOP: Positional Dilution of Precision, a figure of merit used to represent the quality of the satellite geometry when taking GPS readings. PDOPs between 1 and 6 are preferred for most applications, and lower is better.

Perspective convergence: The apparent decrease in inter-object distance as the objects are farther away, for example, the apparent convergence of two railroad rails as they recede into the distance.

Perspective view: A view on a location that includes some relief or perspective distortion, meaning the location of objects may be distorted if their relative distance to the camera varies considerably.

Pixels: Picture elements that make up an image, these are the individual grid cells that record or display a brightness or color in an image.

Plan curvature: Terrain curvature along a contour.

Planar topology: The enforcement of intersection for line and area features in a digital data layer. Each line crossing requires an explicit node and intersection.

Plane surveying: Location surveying methods suitable under the assumption that the surveyed lands form a planar surface, i.e., that distortions due to the Earth's curvature may be ignored.

Platten: The flat back portion of a film camera against which the film rests while an image is collected.

Plumb bob: A weight on a string held freely to determine the local vertical direction.

Pointer: An address stored in a data structure pointing to the next or related data elements. Pointers are used to organize data and speed access.

Polygon: A closed, connected set of lines that define an area.

Polygon inclusion: An area different in some characteristic from the recorded attributes of the polygon, but not resolved.

Position dilution of precision (PDOP): An index of the geometric distribution of a set of satellites for the purposes of estimating and controlling position accuracy. PDOPS typically range between 1 and 20, and lower PDOPs on average result in higher positional accuracies.

Precision: The repeatability of a measure or process.

Primary key: A row or rows in a relational database table that is selected as the key, and that uniquely identifies all the rows in the table.

Prime meridian: See Greenwich meridian.

Profile curvature: Terrain curvature in the direction of steepest descent.

Proximity function: See buffer.

Public Land Survey System (PLSS): A land measurement system used in the western United States of America to unambiguously define parcel location.

Pyramiding: Building images of successively coarser spatial resolution within an image, primarily to allow faster redraws when panning and zooming.

QGIS: An open-source GIS.

Quad-trees: A raster data structure based on successive, adaptive reductions in cell size within a data layer to reduce storage requirements for thematic area data.

Query: Requests or searches for spatial data, typically applied via a database management system.

Radial lens distortion: The displacement of objects in an image due to small lens imperfections, usually radially inward or outward.

Random sample pattern: A sampling pattern where sample location is determined by a random process.

Range distance: A measurement between locations when positioning, usually referring to satellite/receiver distances in GNSS positioning.

Range pole: A pole used in surveying to raise a GNSS antenna, survey prism, or other survey instrument above the ground. Range poles are often used in GNSS data collection to raise an antenna and thereby obtain better PDOPs, and improved accuracy.

Raster data model: A regular "grid cell" approach to defining space. Usually square cells are arranged in rows and columns.

Real time differential correction: GNSS positioning which relies on a radio link and external positioning measurements to correct major GNSS errors in real time, and provide instant improvements in accuracy.

Real time kinematic positioning (RTK): A form of real time differential correction of GNSS positions.

Record: A collection of attributes stored for a specific instance of an entity.

Registration: The conversion or assignment of coordinates from a non-projected coordinate system to a coordinate system, typically via an affine transformation.

Relations: See relational table.

Relational algebra: A set of operations on database tables specified by E.F. Codd for the consistent manipulation of data in a database.

Relational table: A data table in a relational database management system.

Relief displacement: Apparent horizontal distortion of features due to height differences relative to the nadir point in a vertical aerial image.

Remote Sensing: Measuring or recording information about object or phenomena without contacting them.

Resampling: The recalculation and assignment of cell values when changing cell size and/or orientation of a raster grid.

RMSE: Root Mean Square Error, a statistic that measures the difference between true and predicted data values for coordinate locations.

Rubbersheeting: The use of polynomial or other nonlinear transformations to match feature geometry.

Run-length coding: A compression method used to reduce storage requirements for raster data sets. The value and number of sequential occurrences are stored.

Schema: A compact graphical representation of a database conceptual models, entities, and the relationships among them.

Scope: The spatial extent of input for a spatial operation.

Secant lines: Lines of intersection between a developable surface and a spheroid in a map projection.

Selection operation: The identification of a set of objects based on their properties.

Semi-major axis: The larger of the two radial axes that define an ellipsoid.

Semi-minor axis: The smaller of the two radial axes that define an ellipsoid.

Semivariance: The variance between values sampled at a given lag distance apart.

Server: A computer or a program component that stores data, and provides subsets of data in response to requests.

Set algebra: A method for specifying selection criteria based on comparison operators less than, equal to, greater than, and perhaps others.

Shaded relief map: A depiction of the brightness of terrain reflection with a given sun location.

Sidelap: Edge overlap of photographs taken in flightlines.

Singlepart feature: A vector data layer where every feature is individually represented by a point, line, or polygon, and there is a corresponding row in the attribute table for each feature.

Shutter: A system for controlling the time or amount of light reaching a detecting surface.

Skeletonizing: Reducing the width of linear features represented in raster data layers to a single cell.

Sliver: Small, spurious polygons at the margins or boundaries of feature polygons that are an artefact of imprecise digitizing or overlay.

Slope: The change in elevation over a change in location, usually measured over some fixed interval, e.g., the change in height between two points 30 meters apart. Slope is usually reported as a percent slope, or as a degree angle measured from horizontal.

Snap distance: A distance threshold defined in digitizing or other spatial analysis. Point features, nodes, or vertices within the snap are moved to be coincident, to occupy the same location.

Snap tolerance: See snap distance.

Snapping: Automatic line joins during vector digitizing or layer overlay. Nodes or vertices are joined if they are within a specified snap distance.

Spaghetti data model: Vector data model in which lines may cross without intersecting.

Spatial operation: A logical, mathematical, selection, or other spatial process that transforms spatial data.

Spectrum: see electromagnetic spectrum.

Spherical coordinates: A coordinate system based on a sphere. Location on the sphere surface is defined by two angles of rotation in orthogonal planes. The geographic coordinate system of latitude and longitude is the most common example of a spherical coordinate system.

Spheroid: A mathematical model of the shape of the Earth, based on the equation of a sphere.

Spirit leveling: An early leveling survey technique in which horizontal lines were established between survey stations, and relative height differences determined by measured marks on leveling rods.

Spline: A smoothed line or surface created by joining multiple constrained polynomial functions.

SPOT: Systeme Pour l'Observation de la Terre, a satellite imaging system providing 10 to 20 meter resolution images.

SQL: Structured Query Language, a widely adopted set of commands used to manipulate relational data.

SSURGO: Fine resolution digital soil data corresponding to county level soil surveys in the United States. Produced by the Natural Resource Conservation Service.

Standard parallels: Lines of intersection between a developable surface and a spheroid in a map projection.

STATSGO: Coarse resolution digital soil data distributed on a statewide basis for the United States Most often derived from aggregation and generalization of SSURGO data.

State Plane Coordinates: A standardized coordinate system for the United States of America that is based on the Lambert conformal conic and transverse Mercator projections. State plane zones are defined such that projection distortions are maintained to be less than 1 part in 10,000.

Stereo pairs: Overlapping photos taken from different positions but of substantially the same area, with the goal of using parallax to interpret height differences within the overlap area.

Stereographic projection: A map projection with the projection center is placed at the antipode, the point on the opposite side of the spheroid from the projection intersection point with the spheroid.

Stereophotographs: A pair or more of overlapping photographs that allow the perception of three dimensions due to a perspective shift.

Stream mode digitizing: Point data collection via manual digitizing where the distance or time interval between sampled locations is fixed. This removes the need for a button press by the manual operator during digitizing.

Structured Query Language (SQL): A standard syntax for specifying queries to databases.

Survey station: A position occupied, and from which measurements are made, during a land survey.

Systematic sample: A sampling pattern with a regular sampling framework.

Terrestrial reference frame: The set of measured points and their calculated coordinates that are used to define a geodetic datum.

Thematic layer: Thematically distinct spatial data organized in a single layer, e.g., all roads in a study area placed in one thematic layer, all rivers in a different thematic layer.

TIFF: Tagged Image File Format, a widely-supported image distribution format. The Geo-TIFF variant comes with image registration information embedded.

TIGER: Topologically Integrated Geographic Endcoding and Referencing files, a set of structures used to deliver digital vector data and attributes associated with the U.S. Census.

Time-geographic density estimation: A method of estimating the location of a moving object from a time sequence of observation. The method uses time intervals between observations along with estimates of average, maximum, and/or minimum speed to estimate a probable occupation region.

TIN: Triangulated Irregular Network, a data model most commonly used to represent terrain. Elevation points are connected to form triangles in a network.

TM: Thematic Mapper, a high-resolution scanner carried on board later Landsat satellites. Provides information in the visible, near infrared, mid infrared, and thermal portions of the electromagnetic spectrum.

tntMIPS: An image processing and GIS software package produced by Microimages, Inc., of Lincoln, Nebraska.

Topology: Shape-invariant spatial properties of line or area features such as adjacency, contiguity, and connectedness, often recorded in a set of related tables.

Transverse Mercator projection: A common map projection based on a transverse cylinder.

Transaction manager: A component of a database management system that processes requests from clients, and passes them to a server.

Traverse: A series of survey stations spanning a survey. Traverses are closed when they return to the starting point, and open when they do not.

Trigonometric leveling: Measurement of vertical positions or height differences among points by the collection of vertical angles and distance measurements.

Triangulation Survey: Horizontal surveys conducted in a set of interlocking triangles, thereby providing multiple pathways to each survey point. This method provides inherent internal checks on survey measurements.

True North: See Geographic North.

Unary operation: An operation that has only one input.

Undershoot: A digitizing error in which a line end falls short of an intended connection at another line end or segment.

Union: The vertical combination of two spatial data layers, typically over the combined extents of the data layers, and preserving data from both layers.

United States Survey Foot: An official distance used for survey measurements in the United States of America that is slightly different in length from the international definition of a foot.

USGS: United States Geological Survey - the U.S. government agency responsible for most civilian nationwide mapping and spatial data development.

UTM: Universal Transverse Mercator coordinate system, a standard set of map projections developed by the U.S. Military and widely adopted for coordinate specification over regional study areas. A cylindrical projection is specified with a central meridian for each six-degree wide UTM zone.

Variable distance buffer: A buffering variant where the buffer distance depends on some value or level of a feature attribute.

Vector data model: A representation of spatial data based on coordinate location storage for shape-defining points and associated attribute information.

Vertical datum: A reference surface against which vertical heights are measured.

Vertex, vertice: Points used to specify the position and shape of lines.

Virtual reference station (VRS) network: A set of rather closely-spaced GNSS base stations plus communication links that simplify and standardize real-time differential correction within a region.

WAAS: Wide Area Augmentation System, a satellite-based transmission of correction signals to improve GPS positional estimates, largely through the removal of ionsopheric and atmospheric effects.

Wavelength: The distance between peak energy values in an electromagnetic wave.

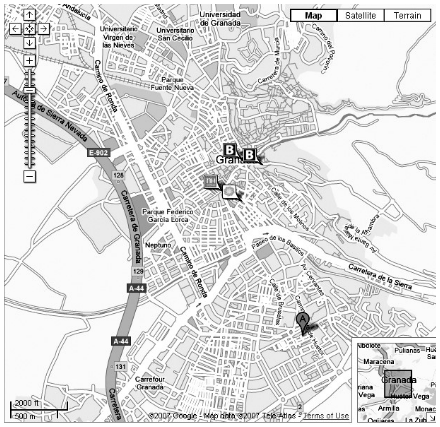
WGS84: World Geodetic System, an Earth-centered reference ellipsoid used for defining spatial locations in three dimensions. Very similar to GRS80 ellipsoid. Commonly used as a basis for map projections.

Zenith angle: The angle measured between a vertical line upward from a point on the Earth and the line from that point to the Sun.

Appendix C: Answers to Selected Study Questions

Chapter 1

1.2: I recently used Google Map (http://maps.google.com/maps?tab=wl) to help plan a trip to Granada, Spain. Data collection consisted of a search with the keywords Granada, Hotel, and Spain. The analysis consisted of the hotel quality ranking, location relative to sites I wanted to visit, and cost. Communication involved sending an image and map to friends.



1.4: GIS software differ from other software primarily in tracking geographic coordinate location, tieing these locations to attribute data, and storing and processing large quantities of data. While many softwares are designed to store and analyze large volumes of data (e.g., video editing), and some other softwares focus on coordinates (e.g., computer assisted design programs for three-dimensional objects), GIS records coordinates that are tied to real, physical locations. Coordinates are defined relative to a physical origin, usually some point on the Earth surface, or the near the center of the Earth, and stored in the computer. Sets of points are combined to characterize the location and shape of geographic features, and non-spatial attributes are associated with these features.

1.6: By our definition in this chapter, paper records and maps are not a GIS, because they are not computer based. However, they do serve in our collection, storage, analysis, and output of spatial data and information, so some would argue that they are a GIS, just an extremely low technology version.

Chapter 2

2.2: Our multiple levels of abstraction from the physical, “real” world usually include a data model, data structures, and machine code. The data model describes the real-world objects with a subset of simple objects and relationships. Data models typically encompass our mental image of how the real world entities are connected, shaped, or related. These models may often be illustrated by box and arrows diagrams. Data structures are how these objects are organized in a computer, for example, what parts go in what files, or how the files are linked on to another. Machine code are the 0’s and 1’s used to store information.

2.4: a) interval/ratio, b) nominal, c) nominal, or ordinal (if read along a brightness gradient), d) ordinal, e) nominal, f) interval/ratio.

2.6: a, b, f

2.8: a) 0.82525625, b) 2.717896524, c) -1.94088466, d) 0.24064227, e) -72.192682, f)128.93670

2.10:

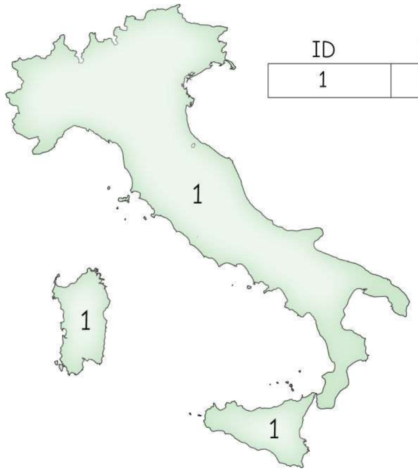
Point	DMS	Decimal Degrees
1	36°45'12"	36.75333
2	114°58'2"	114.9672
3	85°19'7"	85.31816
4	14°00'33"	14.00917
5	275°30'00"	275.00001
6	0°59'43"	0.99528
7	182°19'22"	182.32278

2.12: a) 558.48 km, 4753.72 km, 9523.1 km

2.14: Topology is the study of spatial relationships, and it is important in GIS because topological vector data structures have certain positive properties. Topological relationships such as adjacency, connectivity, proximity, and overlap are important in structuring and analyzing data, and are often helpful in ensuring data quality. Many topological characteristics are invariant to warping or bending spatial features. This is important because we often warp spatial data through map projections.

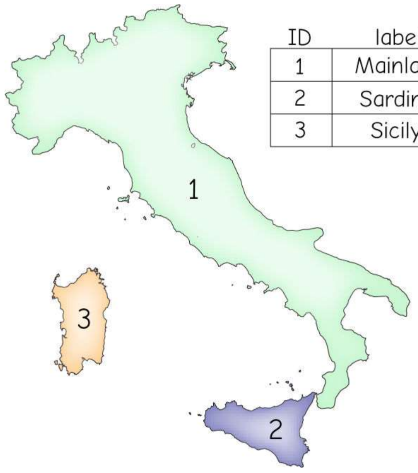
2:16: a) 1, b) 3, c) 2, d) 1, e) 2

2.18: multi-part



ID	label	area
1	Italy	301,340

single-part:



ID	label	area
1	Mainland	251,540
2	Sardinia	24,090
3	Sicily	25,710

2.20:Mixed cells may be a problem under several conditions, for example, when there are very different values for materials within the cell, or when we are interested primarily in one factor that is a minority presence in a cell but not recorded. Mixed cells may be addressed by decreasing the cell size, by carefully developing the assignment rule for cell values when there are mixed constituents, or by recording multiple attributes for each cell, including the identity and proportion of values in each cell.

2.22: Water are E, H, and J.

2.16: a) 14 b) 9 c) 3 d) 133 e) 8 f) 10 g) 145 h) 240

2.24:

2	2	1	2
null	2	2	2
1	null	4	3
3	null	3	4

→

null	1.75
null	3.5

a	a	b	a
a	b	c	b
null	c	c	c
c	c	c	c

→

a	b
null	c

2.26:

One-to-one attribute table
(rows first, start upper-left corner)

cell-ID	count
5	1
7	1
2	1
3	1
3	1
9	1
10	1
4	1
6	1
7	1
8	1
8	1
3	1
4	1
3	1
7	1
7	1
4	1
4	1
3	1
8	1
7	1
4	1
3	1
2	1

One-to-many attribute table

cell-ID	count
2	2
3	6
4	5
5	1
6	1
7	5
8	3
9	1
10	1

2:28: An object data model defines "natural" objects, from the point of view of the model designer, that encompasses spatial and attribute properties, as well as functions or operations that may be specific to that object. Rather that breaking data into thematic layers, components from many themes may exist within an object. Objects may relate to other objects through specific or specialized, unique correspondences or connections.

2.30: a) 1 b) 10111 c) 100000000 d) 100 e)1011 f)1010 g) 11 h) 10100

2.32: a) 4 b) 1 c) 15 d) 43 e) 13 f) 11 g) 129 h) 255

2.34: We compress data when data volumes are too large, particularly for raster data sets. Cells are recorded for each location in a raster area, and gigabytes to terabytes are often stored. Vector data sets typically record shape-defining locations, and only where features of interest occur, for example, a road line. This contrasts with a raster representation which records a set of cells for a road, plus cells for the surrounding area where there is no road.

2.36: Run length codes, by row are:

2:b, 3:a, 1:c, 3:a
2:c, 2:b, 3:d, 2:a
9:b
1:e, 1:c, 1:f, 1:b,1:a 1:d,1:f, 1:b,1:a
1:a, 1:s, 1:a, 3:f, 2:b, 1:a

Chapter 3

3.2 a) 6,372,400 b) 6,356,500 c) 6,344,647

3.4: An ellipsoid is a solid shape based on the rotation of an ellipse. An ellipse is a near circular shape, defined by the equation at right, where x and y are the center of the ellipse and r_1 and r_2 specify how large and flattened the ellipse is. An ellipse becomes a circle when $r_1 = r_2$, and a spheroid is solid based on the rotation of a circle.

Ellipse equation:

$$1 = \frac{(x - x_o)^2}{r_1^2} + \frac{(y - y_o)^2}{r_2^2}$$

3.6: A geoid is usually defined as gravitational equipotential surface chosen as our base for measuring heights. It is an approximately spherical surface for which the force of gravity is a specified constant value. An ellipsoid is a mathematically defined surface, while a geoid is measured, and represents a natural force. The surface of the earth is also a measured surface, but doesn't correspond to a gravitational surface because geologic forces have pushed surface materials above and eroded them below any given equipotential value near the earth surface. We have measured the geoid both from near-surface instruments called gravimeters that measure the gravitational force, and by gravity effects on satellite motion through space.

3.8: Magnetic north is at the point where lines of magnetic attraction converge, and a weightless magnet, if suspended in a frictionless media, would point straight down towards the center of the Earth. Magnetic north is currently located near Greenland.

The geographic north pole is the northern intersection of the Earth's axis of rotation with the Earth's surface. It is located in the Arctic Ocean.

3.10: Multiple datums exist because we have improved datums through time, and because we develop different datums for different purposes. Datums are required for measurements, and so most governments estimated datums when a sufficient number of points were surveyed. Additional points with improved methods will led to subsequent estimations, or versions, of national datums, in most cases with higher accuracies. Satellite and other measurement capabilities developed in the second half of the 20th century added global datums, increasing the number of available datums for most locations.

3.12: Sea level is no longer used as a reference height because sea level is rising, so the mean height would depend on the length of the measurement record and is not stable, there are local variations due to persistent currents, changes in temperature, or changes in salinity, and sea level varies on a 19 year tidal cycle. Technologies have developed such that we can quickly and easily measure accurate height differences smaller than these sources of sea level variability.

3.14:

		NAD27		NAD83(86)		HPGN	
Pnt	State	latitude	longitude	latitude	longitude	latitude	longitude
1	Calif. (S)	32°44'15"	117°09'42"	32°44'15.1827"	117°09'45.1202"	32°44'15.1870"	117°09'45.1201"
3	Wisconsin	43°07'59"	89°20'11"	43°07'58.9806"	89°20'11.4226"	43°07'58.9895"	89°20'11.4192"
5	Colorado	40°00'00"	105°16'01"	40°00'00"	105°16'02.9642"	40°00'00.0068"	105°16'02.9711"
7	Wash. D.C.	38°51'10"	77°02'20"	38°51'10.4052"	77°02'19.9165"	38°51'10.4064"	77°02'19.9041"

3.16:

	NAD83(2011) - 1986		NAD83(2011) - 2015		Surface shift distance (cm)	
Pnt	latitude	longitude	latitude	longitude	latitude	longitude
1	32°44'15"	117°09'42"	32°44'15.0292"	117°09'42.0298"	-90.2	-92.0
3	43°07'59"	89°20'11"	43°07'58.9954"	89°20'10.9973"	14.2	8.4
5	40°00'00"	105°16'01"	40°00'0.0005"	105°16'0.9977"	-1.5	7.1
7	38°51'01"	77°02'21"	38°51'0.992"	77°02'20.9978"	24.7	6.8

3.18:

		NAD27			NAD83(2011)		
Pnt	State	latitude	longitude	elevation (m)	latitude	longitude	elevation (m)
1	Calif.	32°40'00"	-117°00'00"	200	32°40'00.1890"	-117°00'03.0990"	200.654
3	Washington	48°30'00"	-122°00'00"	200	48°29'59.3952"	-122°00'04.5313"	201.149
5	Maine	47°00'00"	-69°00'00"	200	47°00'00.1350"	-68°59'58.1493"	199.887
7	Florida	25°00'00"	-81°30'00"	1	25°00'01.4755"	-81°29'59.3234"	0.582

3.20:

Pnt	State	latitude	longitude	geoid12A elevation (m)	Δheight (cm), to geoid09	Δheight (cm), to geoid99	Δheight (cm), to geoid96
1	Calif.	32°40'00"	-117°00'00"	200	0	5.2	6.6
3	Washington	48°30'00"	-122°00'00"	200	3.2	7.0	11.6
5	Maine	47°00'00"	-69°00'00"	200	1.3	2.7	1.5
7	Florida	25°00'00"	-81°30'00"	1	3.5	6.3	4.5

3.22: MHW = 4.99 above gauge 0. NAVD88 at gauge is 0.43 above gage zero. So MHW is 4.99 - 0.43 above NAVD88, which is 4.56. Hospital should be at 30 + 4.56, or 34.56 ft.

3.24: A developable surface is a mathematical, geometric surface onto which a points are projected from a spheroid or ellipsoid. This developable surface may be mathematically “unrolled” to depict a flat map. Planes, cones, and cylinders are the most common developable surfaces.

3.26: As of December 2015:
Denver, Colorado: NAD83(2011), NAVD88
Latitude and longitudes are 39 45 14.29588(N) 104 53 00.96531(W)

Loma East, California: NAD83(2011), NAVD88
Latitude and longitudes are 32 40 14.00209(N) 117 14 27.75333(W)

Austin CE, Texas: NAD83(2011), NAVD88
Latitude and longitudes are 30 16 48.04361(N) 097 44 16.30349(W)

3.28: The great circle distances are:

Denver to Loma East: 1,359 km

Denver to Austin CE: 1,239 km

Austin CE to Loma East: 1,868 km

3.30: The UTM coordinate system defines map projections for all portions of the globe. Areas between 80° S latitude and 84° N latitude are divided into 6o wide zones, each zone running from the equator to the northern or southern limit. Separate transverse Mercator projections are fit to each zone. Negative zone values are avoided by specifying false eastings and northings, coordinate values added to intermediate projection coordinates.

3.32: Benin, Israel - Transverse Mercator, because both are narrow with the main territorial axes are north-south;
Bhutan - Lambert conformal conic, because they are relatively narrow and have main axes oriented east-west.
Slovenia - either Lambert conformal conic or azimuthal, because although it is has a slight east-west elongation, the shape is somewhat round, and so could be well represented by either forms.

3.34: The Public Land Survey System (PLSS) is a systematic subdivision of land carried out in the U.S. for the purpose of uniquely identifying property boundaries. Principle meridians and baselines are established, and township and range lines surveyed parallel to these at 6 mile intervals. The township/ range grid is further subdivided into 1 mile squares, in turn subdivided into smaller units. The PLSS is not a coordinate system.

Chapter 4

4.2: a, c, and e.

4.4: a) exaggeration, b) simplification, d) simplification, c) omission

4.6: A computer screen is now the most common map media; millions of maps are rendered each hour through applications like google map and mapquest. Paper is the second most common media, and is most used when a hardcopy form is required.

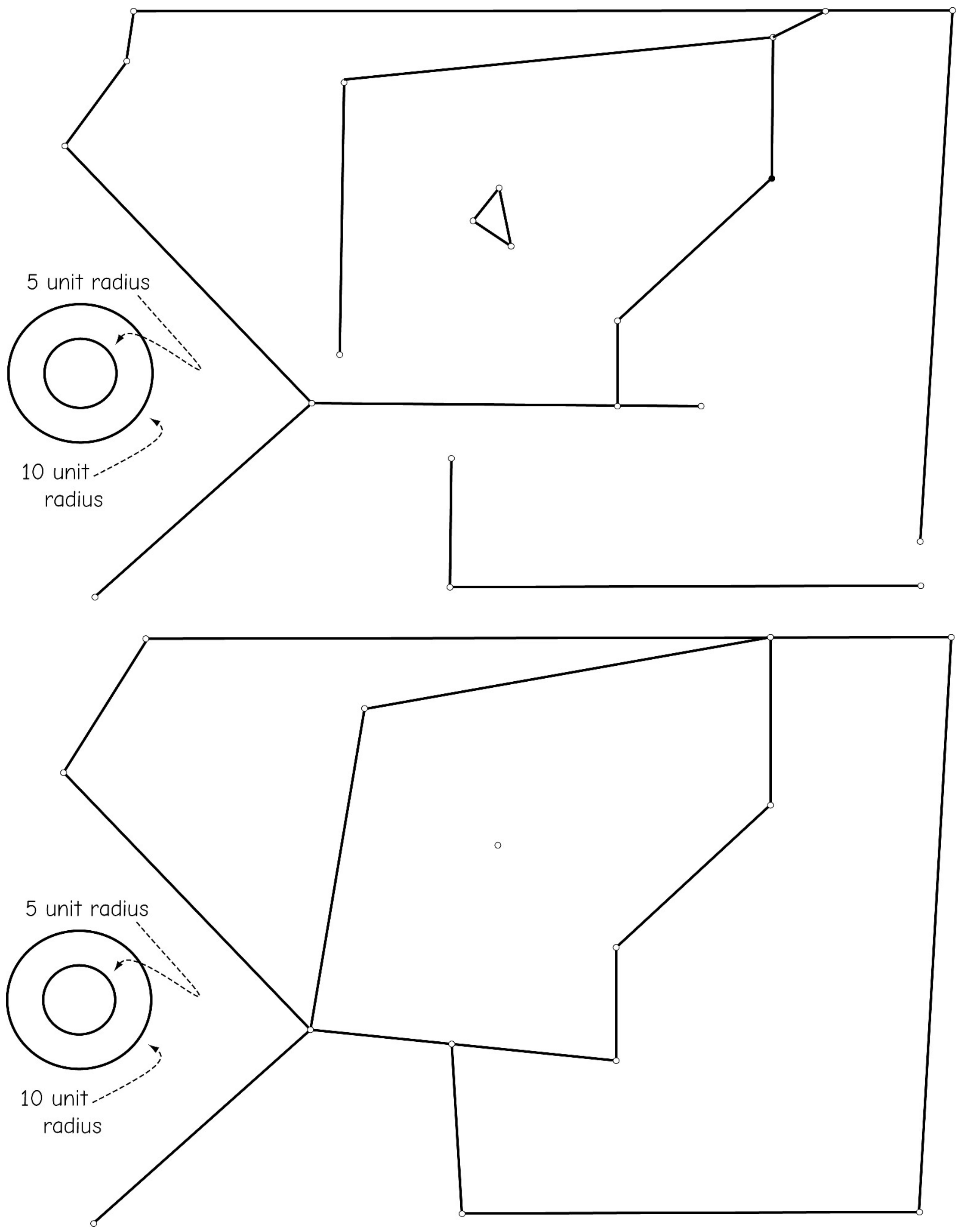
4.8: A large scale map typically shows more detail, because each feature is drawn larger, and there is more opportunity to show variation in shape.

4.10:

Ground distance and units	Corresponding map distance and units	Map Scale
17,120 kilometers	16.85 inches	1 : 40,000,935
23.4 kilometers	11.7 centimeters	1 : 200,000
16.4 miles	9.3 inches	1 : 109,869
102.0 meters	1.855 centimeters	1 : 5,500
320.08 miles	10.24 inches	1 : 2,000,000

4.12: a-undershoot; b-undershoot; c-psuedonode; d-overshoot; e-undershoot; f-missing label; g-overshoot.

4.14: Snap distance 5 (top) and 10 (bottom)



4.16: A spline is a line smoothly fit through a set of points. Splines are used to increase vertex density without substantially slowing the digitizing process, particularly for smoothly-curving features, such as river meanders or winding roads. Splines fit piecewise polynomial functions while imposing smooth join points.

4.18: Manual digitizing involves fixing a map or displaying a scanned image, and manually positioning a pointing device to indicate the location of each node, vertex, or other shape-defining coordinate. Digitized data are in vector form. Scan digitizing uses a machine to record differences in colors or brightness for a map or image document, usually into a raster grid. Lines, points, and areas are defined by some thresholding technique, and lines or points may be thinned and converted to a vector form, as needed. Manual methods have the advantage of low costs for small maps, inexpensive equipment requirements, feature interpretation by humans when using substandard maps, and relatively little training. Scan digitizing is inexpensive for large numbers of very detailed maps, may be automated, and may be more consistent.

4.20: Map registration fixes a map or image to a ground coordinate system so that the coordinates of any point in the media may be determined easily. The process consists of identifying control points that are visible in both the image/map and on the ground, collecting coordinates of these points in both the image/map system and the projected "ground" coordinate system, fitting a system of transformation equations to the coordinate data sets, and applying these transformation equations to the image to convert it to the projected ground coordinate system.

4.22: An affine transformation uses a system of linear equations to estimate the ground easting (E) and northing (N) values from image x and y values. The equations are of the form:

$$E = a_0 + a_1 \cdot x + a_2 \cdot y \qquad N = b_0 + b_1 \cdot x + b_2 \cdot y$$

This is a linear transformation because the x and y variables are not multiplied together or raised to a power larger than 1, by definition the equation of a straight line.

4.24: The average positional error is likely to be the same or larger than the RMSE. The RMSE is usually minimized, or closely related to a minimized quantity when statistically fitting the coordinate transformation. If we collected a representative sample, we expect the RMSE to be approximately equal to the average error. However, if our sampling was inadequate or biased, often it is in areas where we have difficulty identifying good control points, and hence our RMSEs tend to be larger in these locations.

4.24: Transformation b is the most likely to have lowest average error at independently measured points. It depends on the distribution and number of control points, but in most cases higher order polynomials overfit, and while exhibiting lower RMSE values, they have larger errors.

4.28: Metadata are the data about data. They describe the extent, type, coordinate system, lineage, attributes, and other important characteristics of a spatial data set. Metadata are required to evaluate the adequacy of a data set for an intended use.

Chapter 5

5.2: GNSS is based on range distance measurements from multiple satellites to "triangulate" a location. Orbiting satellites transmit radio signals along with precise positioning and timing information. The current distance between a satellite and a receiver is a range measurement. A GNSS receiver combines multiple, simultaneous range measurements to estimate location in near-real time.

5.4: Typically 4 satellites are required for a 3-dimensional fix, although a fix may be determined under some assumptions with data collection from 3 satellites over a short period of time.

5.6: GNSS data range in accuracy, from sub-centimeter for the highest accuracy using carrier phase methods, to tens of meters using real-time C/A positioning. Accuracies are highest when using high quality receiving systems in flat terrain, with no buildings, trees, or other structures to block views of the sky. Accuracies also improve when satellites are widely spaced.

5.8: Figure d depicts the lowest PDOP, with the widest distribution of satellites, closely followed by figure b. Figure a has the highest PDOP, with the tightest distribution.

5.10: Differential positioning is based on the simultaneous measurement of GNSS signals at both a known, base location, and at unknown roving stations. The small errors in range measurement may be calculated for each position measurement at the base station. These range errors may be applied in reverse for corresponding rover data, thereby improving the accuracy of position measurements.

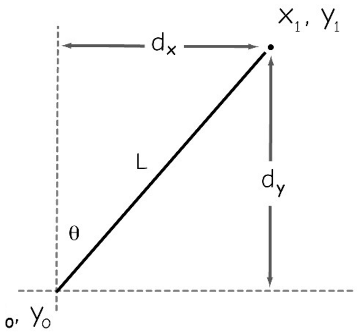
5.12: Dual frequency receivers primarily reduce ionospheric delays, and hence uncertainty in position. They may help somewhat with atmospheric delays, but typically much less than with ionospheric delays.

5.14: GNSS accuracy typically decreases as terrain becomes more varied, or when canopy or buildings obstruct a portion of the sky. Positional accuracy decreases because sub-optimal constellations of satellites are more likely to be observed. Satellites are in closer proximity, and measurements are less independent, and hence to not reinforce each other to improve accuracy.

5.16: WAAS is the Wide Area Augmentation System, a real-time differential correction system designed to aid navigation in U.S. civil aviation. Correction factors are derived from a nationwide network of control stations and broadcast from a geostationary satellite located over the equator. The system is designed primarily for aviation and related uses in North America.

5.18: COGO stands for coordinate geometry. COGO is the calculation of coordinates based on angle and distance measurements.

5.20:



x_0	y_0	θ	L	d_x	d_y	x_1	y_1
0	0	70	100	94.0	34.2	94.0	34.2
15	35	15	130	33.6	125.6	48.6	160.6
400	0	45	200	141.4	141.4	541.4	141.4
150	80	66	20	18.3	8.1	168.3	88.1
10	25	88	12	12.0	0.4	22.0	25.4

5.22:

Point	1	2	3	4	5	6	7
Azimuth	278°	42°	103°	199°	359°	245°	108°14′22″
Bearing	N82°W	N42°E	S77°E	S19°W	N1°W	S65°W	S18°14′22″W

5.24:

Starting point P0, X = 1,200 Y = 400

Point ID	Azimuth	Distance	Delta X	Delta Y	X	Y
P1	95	105	104.6	-9.2	1,304.6	390.8
P2	192	77	-16.0	-75.3	1,288.6	315.5
P3	262	204	-202.0	-28.4	1,086.6	287.1
P4	6	104	10.9	103.4	1,097.5	390.6
P5	18	33	10.2	31.4	1,107.6	422.0
P6	105	88	85.0	-22.8	1,192.6	399.2

5.26

Point (and nearest NGS point)	1 PID DO4877, Cardinal, Minn.	2 PID AB6460 AUS APB3, TX	3 PID DY2143 Sta. Catal. CA	4 PID PX0445 Venus, WY	5 PID DL6000 Chel, Wash.
Latitude	45° 05′ 45.4″	30° 17′ 29.7″	33° 24′ 15.9″	44° 00′ 33.8″	47° 49′ 55.1″
Longitude	93° 00′ 17.9″	97° 41′ 34.3″	118° 24′ 54.2″	109° 30′ 20.2″	119° 59′ 21.6″
Ellipsoidal Height (m)	252.7	141.5	405.9	3604.5	382.1
Geoidal Height (m)	-27.45	-25.95	-36.22	-7.49	-19.27
Orthometric Height (m)	280.15	167.45	442.12	3,611.99	401.37

Chapter 6

6.2: The electromagnetic spectrum is the range of electromagnetic energy frequencies observed. Broadly, this spans from 0 to infinity, however we are most interested in the subset of primary frequencies emitted by the sun. Specifically, we are interested in the ultraviolet through infrared portions of solar radiation, from 0.01 through 1000 µm (1,000,000 µm equals 1 meter). Principal regions of interest are the visible (0.4 to 0.7µm, approximately equally split in the blue, green, and red portions of the spectrum), the near infrared (0.7 to 1.1 µm), and the mid infrared portions (2.5 to 8 µm). Radar wavelengths are important in remote sensing, most often generated from a device, and range from 0.75 cm to 1 m.

6.4: Film is a layered sandwich of emulsions on a polyester base material. The emulsion is sensitive to light, and reacts to darken in a measure proportional to the amount of light (exposure) the layer receives. Different emulsions are sensitive to different spectral regions. Panchromatic film is typically sensitive to visible wavelengths, from 0.4 to 0.7 µm. Color films typically contain three dye layers, sensitive to the blue, green, and red wavelengths (normal color), or green, red, and infrared wavelengths (color infrared film). Spectral reflectance curves plot the sensitivity versus wavelength.

Digital cameras are similar, except that light is typically split by wavelength and directed to separate receptor electronics, one each for each portion of the spectrum observed. Light generates a voltage or current proportional to the light energy, and in this way an image is formed.

6.6: The most common format is the 9-inch mapping camera, in mid to late stages of transition from primarily film-based to primarily electronic sensors. Film is more familiar with nearly 70 years of use and development, and may be less expensive for existing organizations and small projects because the systems are already in hand, and operational. Digital cameras have the advantage of an inherent digital format, obviating subsequent scanning, and may be sharper due to electronic image motion compensation and other image processing. Digital film systems are perhaps more complicated, but also more flexible, and more easily integrated with GNSS, flight control, and other aviation electronics.

6.8: Distortion magnitude varies with mapping cameras, depending on terrain, tilt, camera characteristics, and scale. For vertical photos, typically defined as those with camera axis tilt of less than 3 degrees, errors are typically between 10 and 70 meters over moderate terrain. Errors may be reduced to a meter or less by applying a full photo orthocorrection, a process that analytically removes most tilt and terrain distortion through the three-dimensional geometric analysis and transformation.

6.10: Stereo photographic coverage is the intentional overlap of sequential photographs in a flight line (end lap) and photos in adjacent flight lines (side lap). Overlap provides views of the same set of objections from two different locations. These "perspective views" take advantage of a phenomenon called parallax to reconstruct three-dimensional positions from two-dimensional images.

6.12: Terrain distortion is removed by applying inverse equations that describe the magnitude of terrain-caused distortion. Three dimensional objects that are projected onto a two-dimensional plane are shifted horizontally when they are at different heights. This shift is also dependent on the angle at which the objects are viewed. We may remove the distortion by measuring the height of each point and knowing the viewing angle from the camera location to each point.

6.14: Photointerpretation is the process of converting images into spatial information, typically by an experienced human analyst, or photointerpreter. The photointerpreter uses size, shape, color, brightness, texture, and location to assign or identify characteristics to features of interest.

6.16: The four systems, from Landsat ETM+ through Quickbird represent a range of resolutions (from 30 m through 0.6 m), spectral ranges (full color through near and mid infrared), per scene coverage (from tens of thousands of square kilometers through a few tens of square kilometers), and more the two week to less than two day repeat times. Finally, costs rise markedly along this gradient. Although near the end of its functional life at the time of this writing, the ETM+ data were available for hundreds of dollars for a full scene, while the higher resolution data were from tens to hundreds of thousands of dollars for an equivalent area.

6.18: Image types are selected if they measure the phenomena of interest to the required level of spatial and attribute accuracy, are within the technical capabilities of the organization, have an acceptable probability of successful data collection, and fit with the available budget for acquisition and processing.

Chapter 7

7.2: Do the data provide the required information, for the required area, at the necessary level of categorical and spatial detail, and at the accuracy needed for the intended use?

7.4: Edge-matching is the process of ensuring consistency in features across the edges of mapping projects, areas, and physical maps. When adjacent areas are mapped at different times, by different methods, or by different people there may be incongruent features on either side of the mapping boundary. Roads may not match in location or type, rivers may end abruptly, or the vegetation or elevation change in an impossible manner. Edge-matching attempts to resolve these differences, and if possible, remove errors across mapping boundaries.

Chapter 8

8.2: Database management systems are computer software tools that aid in the entry, organization, analysis, distribution, and presentation of data.

8.4: A one-to-one relationship among table means that for every row in one table that in some way is matched to a row in another table, there is only one row in the second table that matches. A many-to-one relationship means that one row in a table may match many rows in a second table. Note that by match, we do not mean completely match. Usually we are using a column in each table to match the tables; the rows are considered to match when the match column has the same value in both tables.

8.6: Osel, NumT

8.8: The eight basic operations are illustrated in the section "Primary Operations" in Chapter 8. They are restrict, project, product, divide, union, intersect, difference, and join.

8.10: Sets from OR conditions will have the same number or more members than the component conditions.

8.12: a) Florida, Georgia, Iowa, Minnesota, Wisconsin.
b) Iowa, Minnesota, Wisconsin.
c) Alabama, Alaska, Florida.
d) Minnesota, Alaska.
e) Iowa, Oklahoma.
f) Minnesota, Wisconsin.

8:14: Normal forms are a way of organizing database tables. When followed, they optimally structure tables to remove redundancies, efficiently store data, and organize data in "natural" groupings that speed analysis and increase flexibility.

8.16:

Id1	pos
Y	wa
Z	ea
A	rt
Y	pr
Y	nn
R	rt
Q	mn

Id2	tm
X	5
Y	1
A	6
Q	4
N	3
L	2

→

Id1	pos	tm
Y	wa	1
Z	ea	-
A	rt	6
Y	pr	1
Y	nn	1
R	rt	-
Q	mn	4

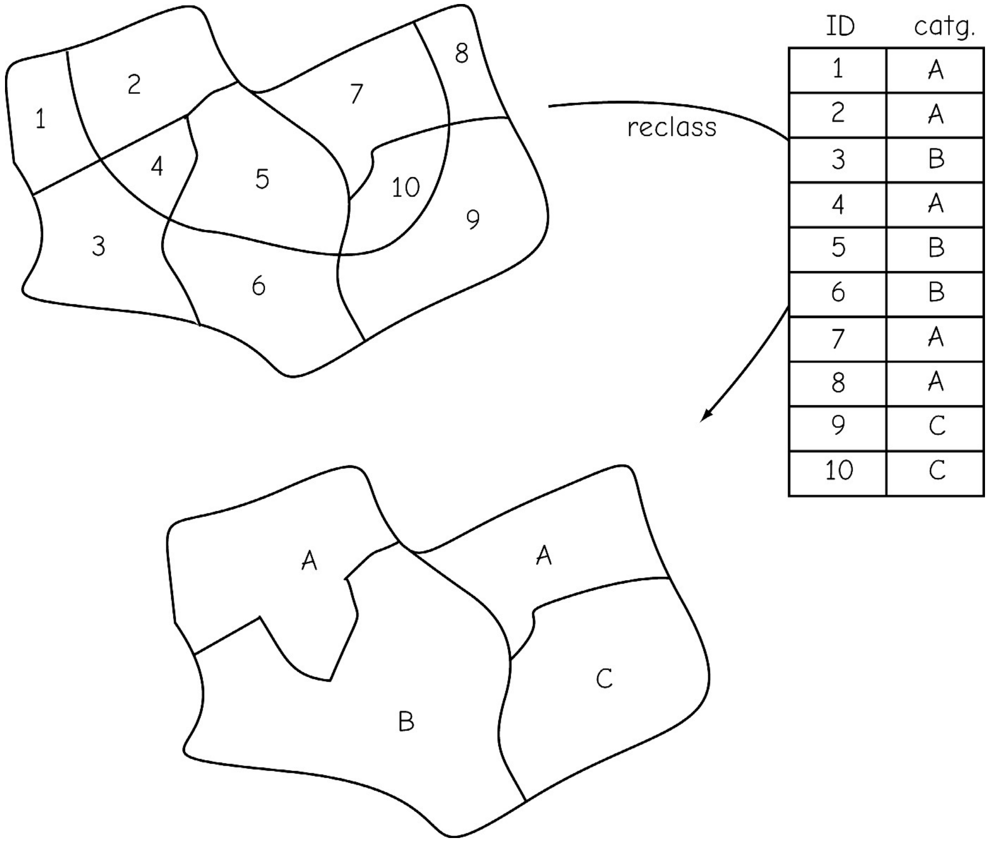
8.18: A functional dependency occurs when knowing the value of one variable defines the value of another variable. For example, if I know a person that a person was a German citizen in 2014, then I know that their Chancellor was Angela Merkel, or if that a person is a Chicago Cubs fan in 2015, then their team has the longest active period without winning the baseball World Series.

8.20: ID -> Size, Color;
Size -> Color
Source -> ID, Size, Shape, Color, Age

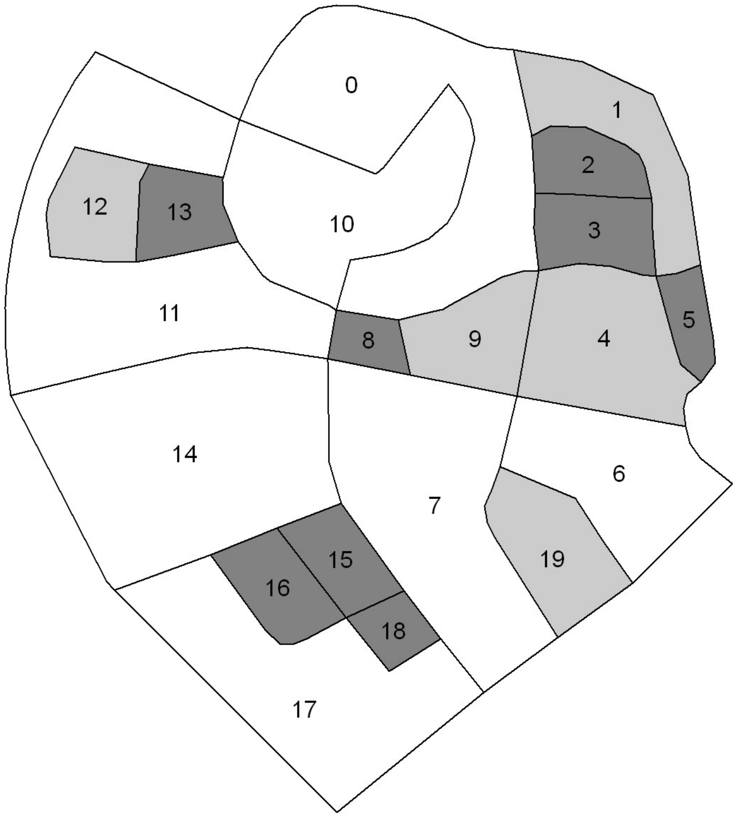
Chapter 9

9.2: Selection operations apply criteria to features, and identify features that meet those criteria. The criteria may apply to spatial characteristics, for example, the size, shape, or location of a polygon; they may apply to non-spatial attributes of the features, for example the value or condition of an attribute.

9.4: a) B or C; b) A and B; c) [A and B] and not C; d) [B or C] and not [B and C]

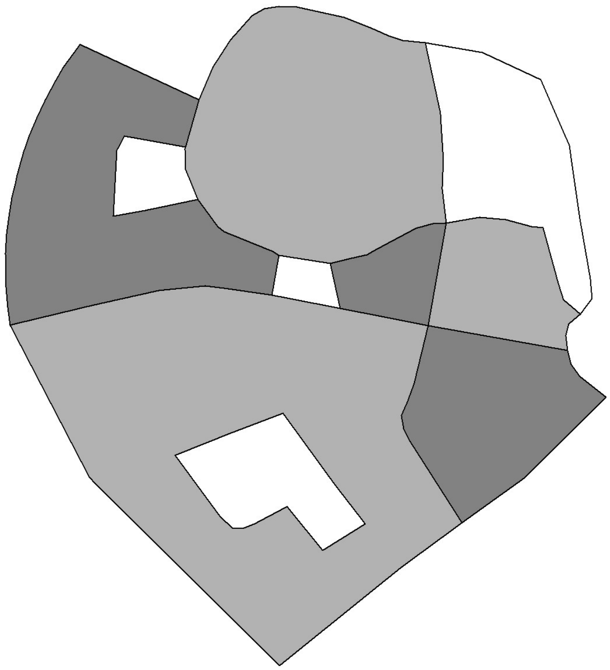


9.8: Large are white, medium light gray, small darker grey.

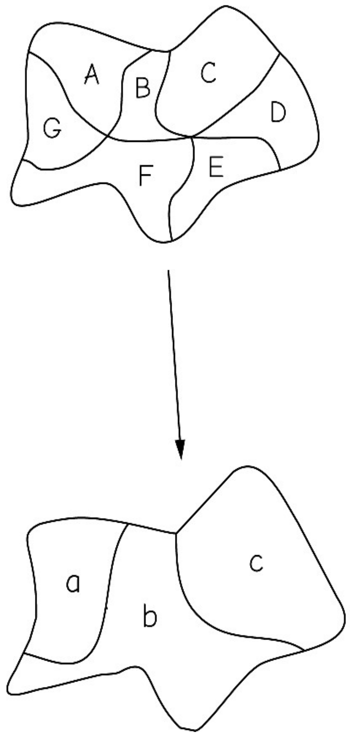


9.10: The modifiable area unit problem arises because statistics for aggregated areas depend on the aggregations. We may combine adjacent areas, and calculate sums, means, medians, and other attributes of the areal units. If we are selective about how we aggregate, we may change these statistics solely by changing the aggregation units.This is the modifiable areal unit problem. The zoning effect is how aggregate statistics change with zone boundaries. The area effect is how statistics change when changing the size of aggregation areas.

9.12



9:14:

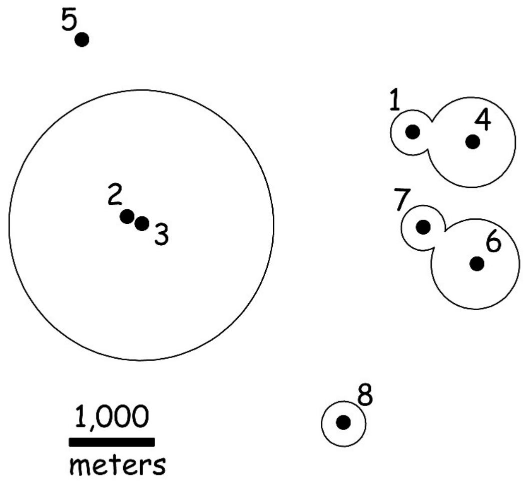


ID	Class	Count	Type	Cost
A	1	11	Farm	1,000
B	2	9	Farm	900
C	1	3	Ranch	1,100
D	1	4	Suburb	200
E	2	21	Suburb	700
F	2	14	Farm	800
G	1	6	Ranch	1,200

NewID	Class	Count
a	1	17
b	2	44
c	1	7

9.16: Multi-distance, retain, exterior.

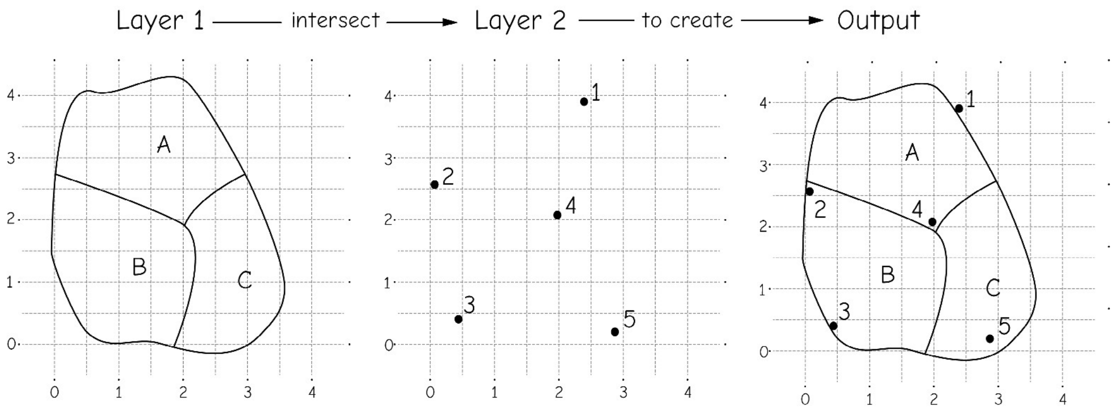
9.18:



ID	distance
1	250
2	0
3	1500
4	500
5	0
6	500
7	250
8	250

9.20: The minimum dimension (point, line, or polygon) is chosen because to do otherwise courts ambiguity. If two lower dimension features are coincident with higher-dimension features, it is unclear how the attributes should be recorded in the resultant features. For example, if two points fall within a polygon, the polygon attributes may be unambiguously associated with each point. It is unclear or at best cumbersome to assign both sets of point attributes to an output polygon.

9.22:



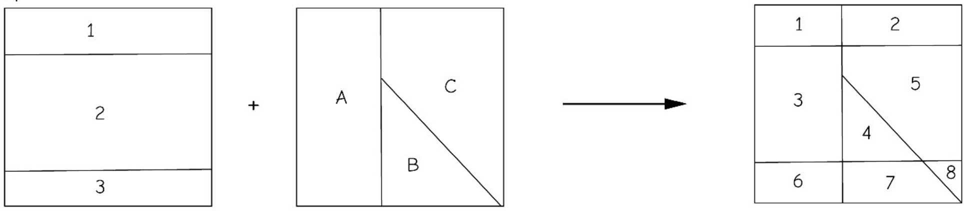
ID	code	stts
A	Grs	fee
B	Wtl	lease
C	Urb	TxDf

ID	ctgry
1	FtTyr
2	XC
3	STrk
4	Exp
5	CFbr

ID	ctgry	code	stts
1	FtTyr	null	null
2	XC	Grs	fee
3	STrk	Wtl	lease
4	Exp	Grs	fee
5	CFbr	Urb	TxDf

9:24:

Spatial Data



Tables

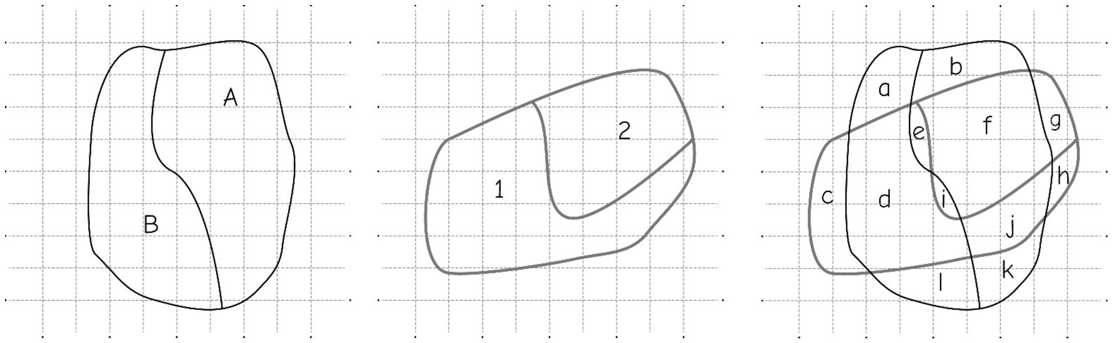
ID	no	own
1	1215	E. Nuff
2	1863	U. Kahn
3	1919	I. Cee

+

ID	Ward	Rep	pop.
A	11	Hanks	496
B	2	Ryan	187
C	9	Gump	2,312

ID	no	own	Ward	Rep	pop.
1	1215	E. Nuff	11	Hanks	496
2	1215	E. Nuff	9	Gump	2,312
3	1863	U. Kahn	11	Hanks	496
4	1863	U. Kahn	2	Ryan	187
5	1863	U. Kahn	9	Gump	2,312
6	1919	I. Cee	11	Hanks	496
7	1919	I. Cee	2	Ryan	187
8	1919	I. Cee	9	Gump	2,312

9:26:



ID	Lndcov	county
A	forest	Wallowa
B	urban	Lincoln

ID	type	txtr
1	Evern	sand
2	Gillard	clay

ID	Lndcov	county	type	txtr
a	urban	Lincoln	null	null
b	forest	Wallowa	null	null
c	null	null	Evern	sand
d	urban	Lincoln	Evern	sand
e	forest	Wallowa	Evern	sand
f	forest	Wallowa	Gillard	clay
g	null	null	Gillard	clay
h	null	null	Evern	sand
i	urban	Lincoln	Gillard	clay
j	forest	Wallowa	Evern	sand
k	forest	Wallowa	null	null
l	urban	Lincoln	null	null

9:28:

Layer 1

A	A	A	B
A	A	B	B
A	C	C	B
A	C	B	B

Layer 2

X	X	X	X
W	W	X	X
W	W	W	X
W	W	W	W

→

1	1	1	2
3	3	2	2
3	5	5	2
3	5	4	4

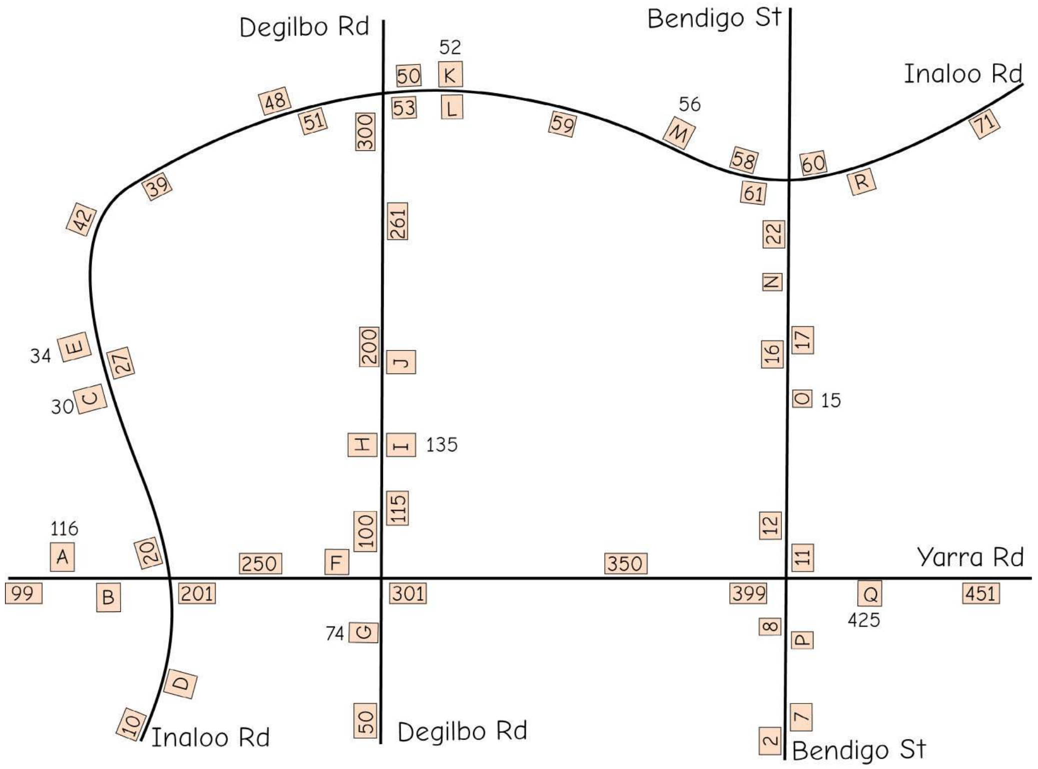
ID	type	count
1	A	7
2	B	6
3	C	3

ID	type	count
1	X	7
2	W	9

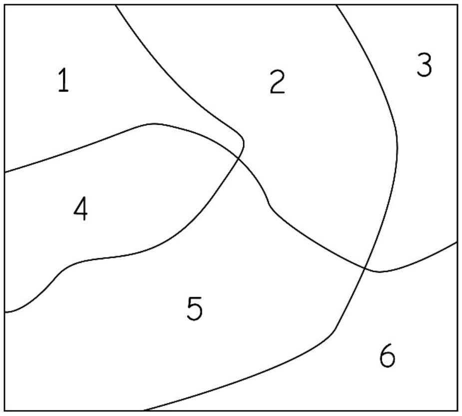
ID	T1	T2	count
1	A	X	3
2	B	X	3
3	A	W	4
4	B	W	2
5	C	W	3

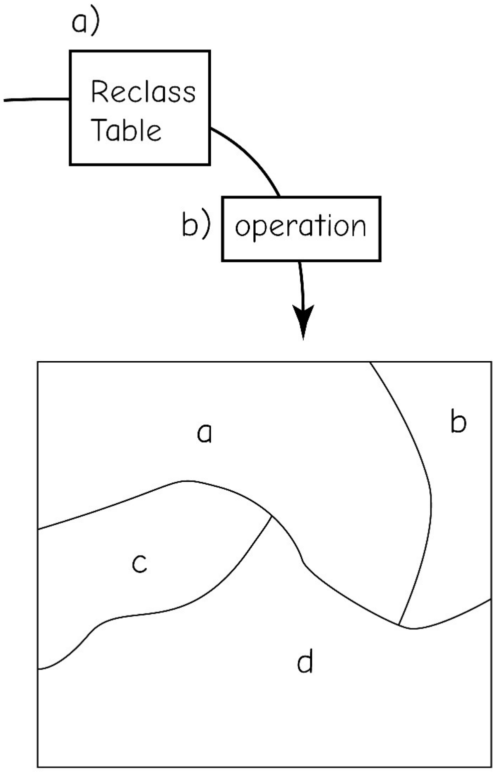
9.30: Network models are connected linear graphs through which resources flow, or to which movement may be constrained. There may be both source and demand features connected to these networks. Networks are different from many other spatial models in that movement or occurrence is limited to the network, and they often track time-varying

9:32: Note answer values in cases large address ranges may be off by one address unit



9.34:





a)

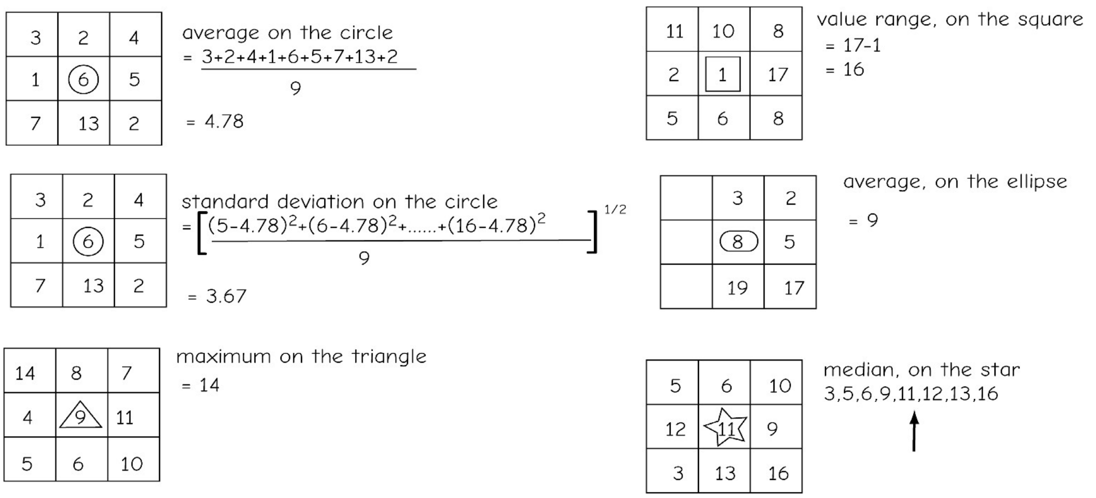
in	out
1	a
2	a
3	b
4	c
5	d
6	d

b) dissolve

Chapter 10

10.2: Compatible cell sizes are often required in raster operations because otherwise the input is often ambiguous. If one input cell is substantially larger or mismatched to another, it may be uncertain which input cell value to choose.

10.4:



10.6: C1 = 1, C2 = null, C3 = 0, C12 = 0.

10.8: C1 = 1, C3 = 1, C4 = 0, C10 = 4

10.10: Any nested operation, something like con(isnull(layer1, layer2, layer1), or sqrt(abs(layer1)).

10.12: C7 = null, C8 = null, C13 = 1, C16 = null

10.14: In most systems, a NULL value is returned when it appears as any input of an operation, unless there are explicit instructions to ignore null values.

10.16:

1	1	0	0
0	0	0	1
1	1	0	0
1	0	0	1

and

0	1	0	1
1	0	0	1
1	1	1	0
1	0	1	1

=

0	1	0	0
0	0	0	1
1	1	0	0
1	0	0	1

1	1	0	0
0	0	0	1
1	1	0	0
1	0	0	1

or

0	1	0	1
1	0	0	1
1	1	1	0
1	0	1	1

=

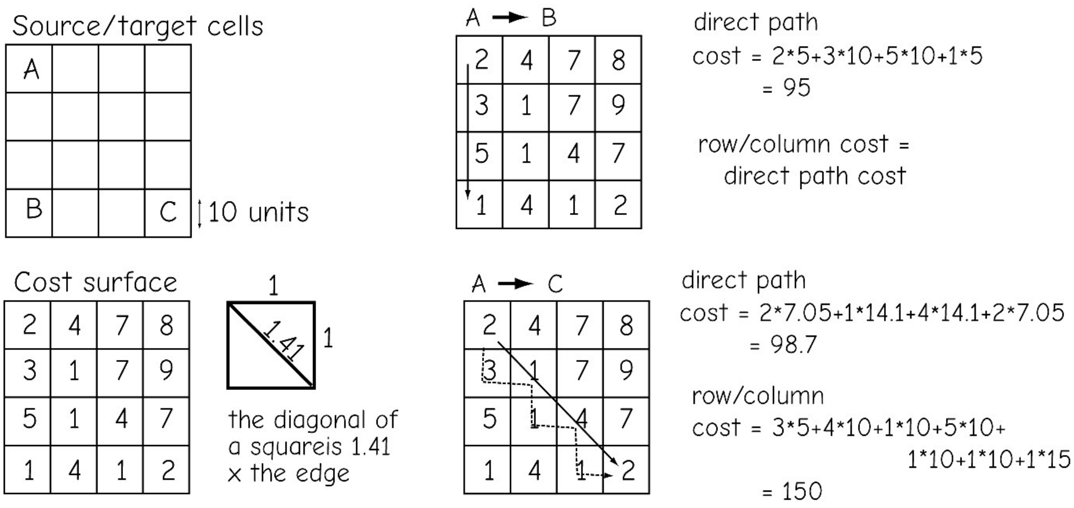
1	1	0	1
1	0	0	1
1	1	1	0
1	0	1	1

10.18: A clip function may be implemented in a raster environment by creating a clip layer which have cells with a value of 1 wherever data in the target layer are to be kept, and 0 cell values corresponding to areas where the target layer data is to be discarded.

10.20: The kernel is a high-pass filter. It would highlight local differences in the target raster, with values changed little in areas where cell values are about the same, and exaggerating values that are higher or lower than their neighbors.

10.22: High spatial covariance means cells near each other tend to have similar values; low values tend to be clustered near low values, and high values clustered near other high values.

10.24:


Source/target cells
A
B
C
10 units
Cost surface
2 4 7 8
3 1 7 9
5 1 4 7
1 4 1 2
1
1.41
1
the diagonal of
a squareis 1.41
x the edge
A → B
2 4 7 8
3 1 7 9
5 1 4 7
1 4 1 2
direct path
cost = 2*5+3*10+5*10+1*5
= 95
row/column cost =
direct path cost
A → C
2 4 7 8
3 1 7 9
5 1 4 7
1 4 1 2
direct path
cost = 2*7.05+1*14.1+4*14.1+2*7.05
= 98.7
row/column
cost = 3*5+4*10+1*10+5*10+
1*10+1*10+1*15
= 150

Chapter 11

11.2: Digital elevation models are created by a variety of methods. Leveling or other ground surveys are used to measure relative height differences across profiles, using optical and electronic instruments to measure distance and vertical and horizontal angles. Photo-based methods rely on parallax, the relative displacement of objects depending on their distance from an observation point. Downward looking images taken from aircraft or satellites may be combined with knowledge of aircraft position and ground surveys to create DEMs. Laser and radar measurements from airborne platforms are a third common method for DEM creation. Return times are recorded for electromagnetic signals sent form the aircraft, and used to calculate terrain height relative to the aircraft. These are combined with precise positioning information for the aircraft and with previous ground surveys to produce accurate DEMs.

11.4:

windows | 4-nearest cell | 3rd-order finite difference

a)

110	113	118
112	114	119
111	117	121

←10→

4-nearest cell:

$$dz/dx = \frac{119 - 112}{20} = 0.35$$

$$dz/dy = \frac{113 - 117}{20} = -0.20$$

3rd-order finite difference:

$$dz/dx = \frac{118 + 2*119 + 121 + -110 - 2*112 - 111}{80} = 0.40$$

$$dz/dy = \frac{118 + 2*113 + 110 + -121 - 2*117 - 111}{80} = -0.15$$

b)

67	63	62
65	64	64
70	68	66

4-nearest cell:

$$dz/dx = \frac{64 - 65}{20} = -0.05$$

$$dz/dy = \frac{63 - 68}{20} = -0.25$$

3rd-order finite difference:

$$dz/dx = \frac{62 + 2*64 + 66 + -67 - 2*65 - 70}{80} = -0.14$$

$$dz/dy = \frac{62 + 2*63 + 67 + -66 - 2*68 - 70}{80} = -0.21$$

c)

18	23	17
21	24	19
20	22	18

4-nearest cell:

$$dz/dx = \frac{19 - 21}{20} = -0.1$$

$$dz/dy = \frac{23 - 22}{20} = 0.05$$

3rd-order finite difference:

$$dz/dx = \frac{17 + 2*19 + 18 + -18 - 2*21 - 20}{80} = -0.09$$

$$dz/dy = \frac{17 + 2*23 + 18 + -18 - 2*22 - 20}{80} = -0.01$$

11.6: Slope and aspect, four nearest neighbor method.

712	709	707	703	704
710	a 706	704	700	696
708	705	705	c 697	700
711	b 709	705	696	694
714	712	708	703	698

a) dz/dx = -0.30, dz/dy = 0.20,
slope = 19.83°,
aspect = 123.69°

b) dz/dx = -0.30, dz/dy = -0.35,
slope = 24.75°,
aspect = 40.6°

c) dz/dx = -0.25, dz/dy = 0.2,
slope = 17.75°,
aspect = 128.66°

11.8: Slope and aspect, third-order finite difference method.

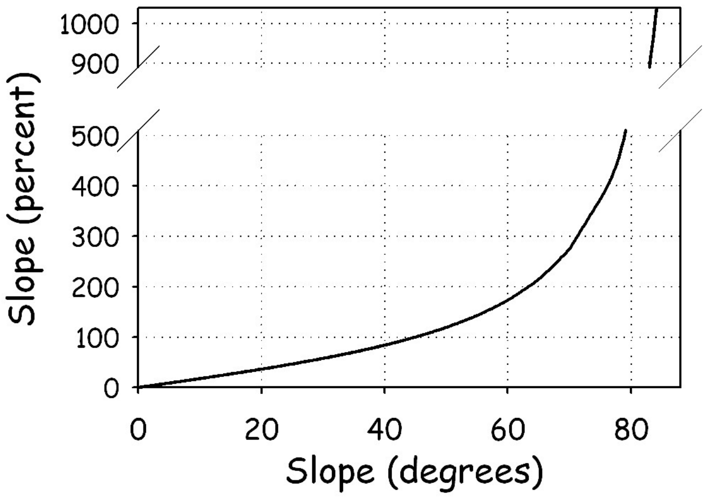
712	709	707	703	704
710	a 706	704	700	696
708	705	705	c 697	700
711	b 709	705	696	694
714	712	708	703	698

a) dz/dx = -0.25, dz/dy = 0.18,
slope = 16.97°,
aspect = 124.0°

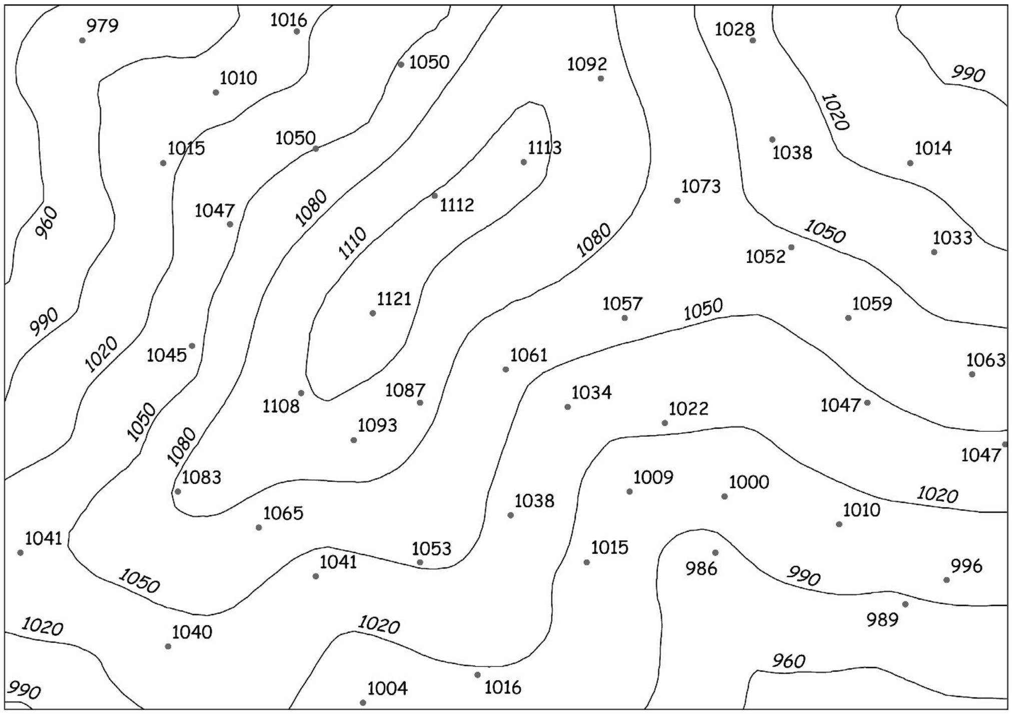
b) dz/dx = -0.26, dz/dy = -0.29,
slope = 21.27°,
aspect = 42.4°

c) dz/dx = -0.36, dz/dy = 0.11,
slope = 20.78°,
aspect = 107.24°

11.10: Slope as percent is larger over most of the possible range.



11.12

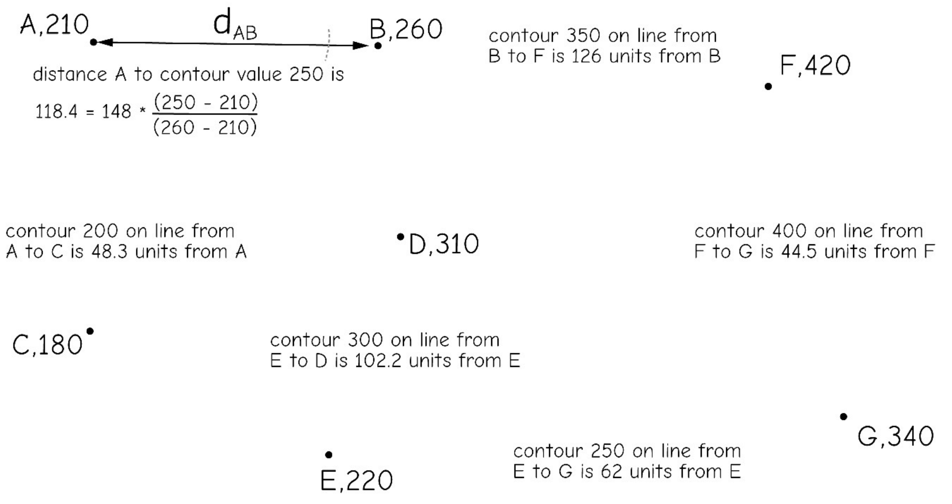


11.14: The formula for contour calculations is:

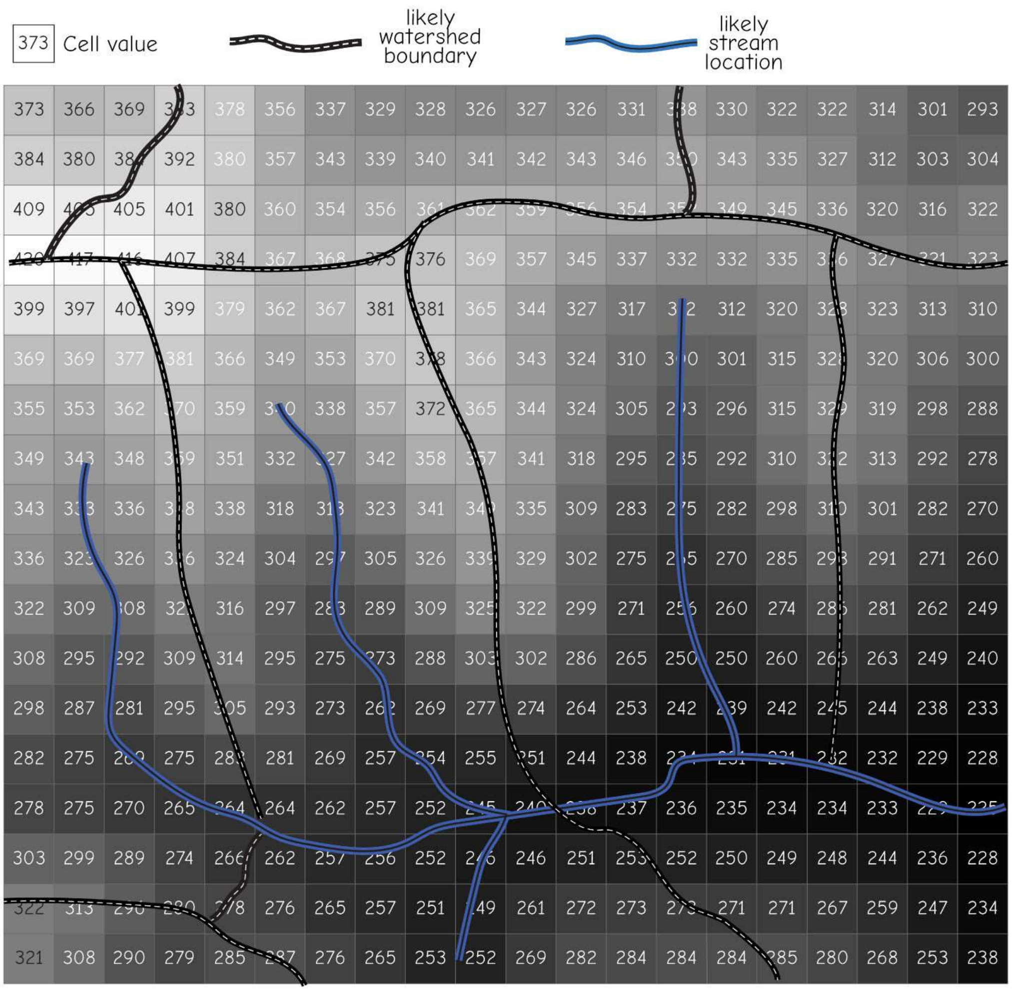
$$d_2 = d_1 \cdot \frac{H_t - H_c}{H_t - H_b}$$

where d_2 is the distance from the upper point to the point on the contour, H_t and H_b are the upper and lower elevations at known points, and d_1 is the distance between points H_t and H_b.

11.16:


A,210
d_{AB}
B,260
distance A to contour value 250 is
118.4 = 148 * (250 - 210) / (260 - 210)
contour 350 on line from B to F is 126 units from B
F,420
contour 200 on line from A to C is 48.3 units from A
D,310
contour 400 on line from F to G is 44.5 units from F
C,180
contour 300 on line from E to D is 102.2 units from E
G,340
E,220
contour 250 on line from E to G is 62 units from E

11.18


373 Cell value
likely watershed boundary
likely stream location

11.20: A solar zenith angle is the angle measured at an observation point between the vertical line, “straight up,” and the sun’s location. The solar azimuth angle is the angle turned clockwise from geographic north to the sun’s location. The solar incidence angle is the angle between an incoming sun’s ray and the surface normal line.

11.22: A viewshed for a point is the combination of areas that are visible from a point. Viewsheds are used in landscape design to maximize scenic vistas or hide powerlines, roads, or other features, and they are used in telecommunications to calculate inter-visibility and communication networks. They are calculated by tracing rays from a view point to all viewable locations, using a DEM to estimate if the angles to all intervening points are lower than the angle to the potentially visible point.

Chapter 12

12.2: The four discussed sampling patters are:

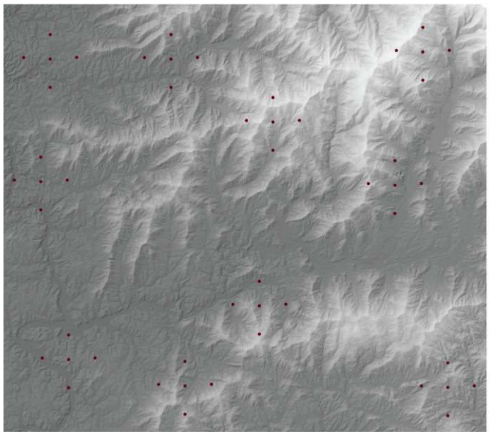
1) random, where points are assigned x and y locations randomly drawn from those in the area of interest,

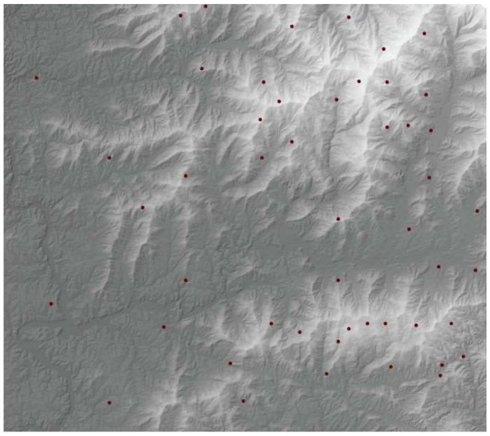
2) systematic, where locations are assigned in a fixed patter, e.g., spaced every 100 meters in the x and y direction from a starting point,

3) clustered, where points are assigned in a systematic or restricted random manner from each of a set of starting points, and

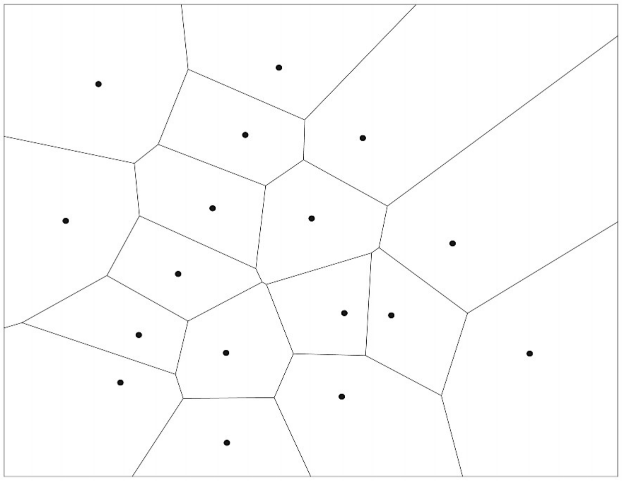
4) adaptive, where local sampling density is related to variability, with more samples in more variable areas

12.4: Adaptive sampling will likely give a better estimate of the surface over all numbers of sampling points.



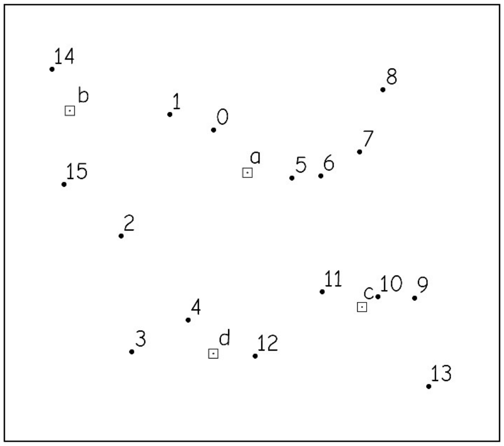


12.6:



12.8: a is 18; b is 12; c is 6; d is 10.25; e is 0; f is 22.

12.10:



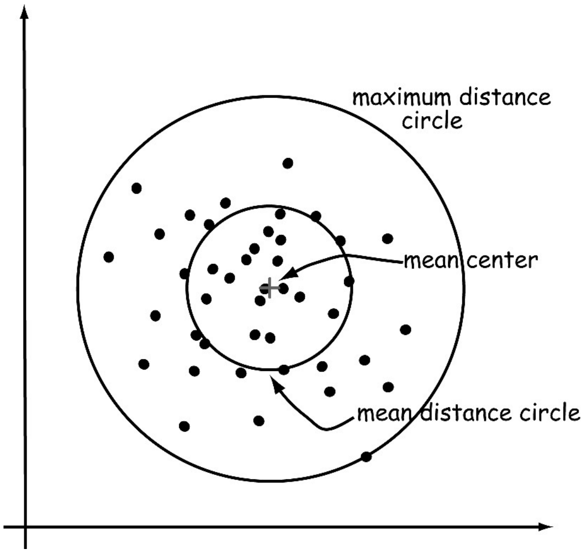
Unknown points:

ID	X	Y	Z
a	155,859.6	4,477,159.0	2,200.95
b	147,580.6	4,478,884.2	1,872.06
c	162,535.7	4,469,960.2	2,059.26
d	151,714.9	4,463,755.2	2,288.4

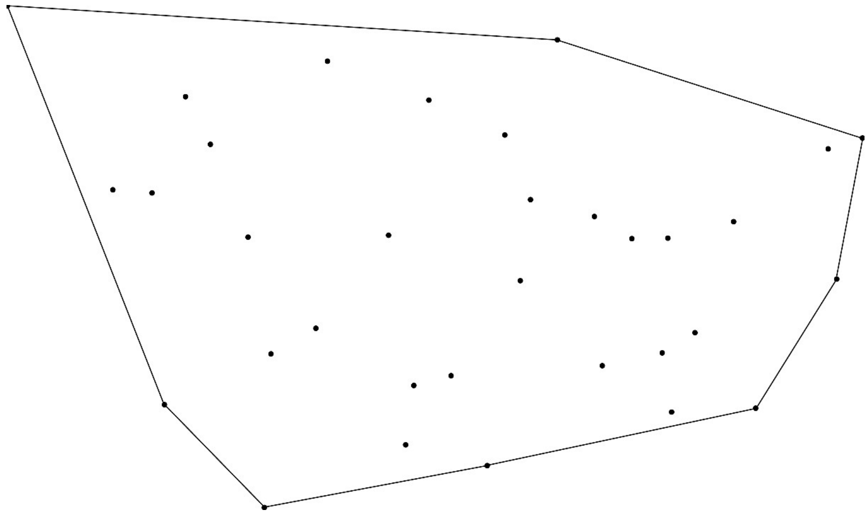
ID	X	Y	Z
0	153,951.9	4,478,714.6	2040.6
1	151,280.9	4,479,647.3	1863.0
2	148,228.5	4,472,143.4	1992.1
3	148,906.8	4,464,978.6	2540.1
4	152,383.2	4,466,928.8	2106.3
5	158,827.3	4,475,746.9	2283.2
6	160,607.9	4,475,874.2	1933.5
7	163,024.4	4,477,357.9	1836.4
8	164,465.8	4,481,173.5	1838.3
9	166,416.0	4,468,285.4	2523.9
10	164,169.1	4,468,370.4	2138.8
11	160,692.7	4,468,709.3	1854.2
12	156,537.9	4,464,724.2	1866.9
13	167,306.3	4,462,816.5	2453.8
14	143,946.6	4,482,445.4	1837.9
15	144,709.7	4,475,323.0	1912.8

12.12: A trend surface interpolator estimates coefficients for an equation globally - all data are used to estimate the coefficients for a prediction equation, and apply across the entire sample region. A kriging interpolator uses estimates of global and local variation, specifically spatial autocorrelation, to estimate coefficients for prediction equations. In this way, samples can influence predictions depending on the observed spatial autocorrelation and distribution of samples.

12.14:



12.16:



12.18: Kernel mapping assigns a density surface around observation points. The density surface represents the likelihood or probability that the organism or population occupied the area. “Stacking,” or adding together the density surfaces for a set of observations creates a combined probability map, giving an estimate of the occupancy across combined observations or samples. The density surface is typically described mathematically by an equation, with the shape controlled by parameters. There is often a “bandwidth” parameter that affects the “peakedness” of each individual density surface.

12.20: B shows the wider bandwidth, because the distributions are broader around each point, and the maximum values are lower, indicated by less saturated shading near the point locations.

Chapter 13

13.2: Criteria often must be refined because they are expressed in a way that may not be directly applied. Distances or groupings may be ambiguously described, e.g., near, far, large, or small, and often these must be quantified before entry into a cartographic model.

13.4: A discrete weighting has specific, distinct categories. Roads may be passable or impassable for large vehicles, rivers deep or shallow, or forests evergreen or conifer dominated. Weightings may be defined which are specific to these categories, e.g., passable roads receive a weight of 0.5, while impassable roads a weight of zero. In contrast, weightings may also be continuous, in that each road may have a measured width, e.g., 12.4 meters, and the weight may be some continuous function of this width, e.g., width *132.7.

13.6: A - original
B - reclassed by T, with low and high input values reclassed to high output values, and intermediate input values reclassed to low output values;
C - reclassed by W, with low input values set to zero when below a threshold, and then a linear increase in output with input above the threshold;
D - Reclassed by R. High values above a threshold are set to zero, and low values below the threshold are set to 1.

13.8: Flowchart D is the most plausible
B - Clip of high density by park buffer incorrectly provides the complement of the distance from current park criteria, selecting the wrong areas; Unions school buffer and HDNP buffer, would also need a subsequent selection to extract appropriate areas;
C - Unions high density and park buffer. Would need a select and delete, or a clip or some other operation to correctly apply the distance from part criteria.

13.10:Correct answer is C
A - applies area test too early, later intersections may split candidate polygons, rendering smaller area polygons than the size limit specified for the analysis;
B - commits same error, later in the process;
D - omits raster to vector conversion on elevation, and allows areas in same parcel that are satisfactory and adjacent, but split by other criteria to be in different polygons, so may reject acceptable areas.

Chapter 14

14.2: Data transfer is perhaps the process most commonly aided by spatial data standards. Among other things, standards require that spatial data sources, methods, and characteristics be described in a consistent manner. Spatial data standards provide a predictable way of organizing data so that it may be transferred among organizations. Standards allow data to be transferred without loss of information.

14.4: The mean reports the central error, the average error one would expect when sampling from a population. A frequency threshold describes the percentage of errors above or below a value. This gives some general notion of the likelihood of large or small errors.

14.6: Positional accuracy reports how close the represented locations of objects are near the true locations of the objects. Attribute accuracy reflects how often the value of a categorical attribute is correct (discrete) or how close and interval/ratio attribute is to the true value (continuous). Logical consistency does not imply either spatial or attribute accuracy, but just that multiple themes or types of are consistent, e.g., there are no roads in a lake, fires on salt flats, or oil deposits in granitic rocks.

14.8: The steps in applying the NSSDA are 1) identify test points, 2) identify source for “true” points, and extract truth corresponding to test points, 3) Calculate the positional error for each true/test pair, 4) record the error data in a standardized table, which includes calculation of error statistics, 5) create the documentation/metadata describing the accuracy assessment.

14.10: Good candidate points are any features that are well-defined and may be visible on both the data set to be tested and in the source used for truth. This often means constructed features, for example road intersections, curbs, manhole covers, geodetic markers, utility poles, or fire hydrants or other relatively immobile features.

14.12: Metadata are the “data about data.” They are important because the describe the characteristics about any data we might wish to use. The allow us to evaluate the data suitability for intended uses, maintain the investment over multiple organizations or changes in personnel, and help us in explaining or describing our data to others.